
**UNIVERSIDADE FEDERAL DE PERNAMBUCO**

**CENTRO DE TECNOLOGIA E GEOCIÊNCIAS**

**DEPARTAMENTO DE ENGENHARIA BIOMÉDICA**

**WHENDEL MUNIZ DOS SANTOS**

# CLASSIFICAÇÃO DE TIPOS DE CÂNCER EM MULHERES A PARTIR DE APRENDIZADO DE MÁQUINA E SEQUENCIAMENTOS GENÉTICOS

**RECIFE**

**2024**

**UNIVERSIDADE FEDERAL DE PERNAMBUCO**

**CENTRO DE TECNOLOGIA E GEOCIÊNCIAS**

**DEPARTAMENTO DE ENGENHARIA BIOMÉDICA**

**WHENDEL MUNIZ DOS SANTOS**

# CLASSIFICAÇÃO DE TIPOS DE CÂNCER EM MULHERES A PARTIR DE APRENDIZADO DE MÁQUINA E SEQUENCIAMENTOS GENÉTICOS

Trabalho Supervisionado, apresentado ao Curso de Engenharia Biomédica do Centro de Tecnologia e Geociências da Universidade Federal de Pernambuco, como requisito para a obtenção do título de Bacharel em Engenharia Biomédica.

**Orientador(a):** Prof. Dr. Wellington Pinheiro dos Santos
**Coorientador(a):** Profa. Dra. Maíra Araújo de Santana

**RECIFE**

**2024**

Ficha de identificação da obra elaborada pelo autor,
através do programa de geração automática do SIB/UFPE

Santos, Whendel Muniz dos .
Classificação de tipos de câncer em mulheres a partir de aprendizado de máquina e sequenciamentos genéticos / Whendel Muniz dos Santos. - Recife, 2024.
47 p. : il., tab.

Orientador(a): Wellington Pinheiro dos Santos
Cooorientador(a): Maíra Araújo de Santana
Trabalho de Conclusão de Curso (Graduação) - Universidade Federal de Pernambuco, Centro de Tecnologia e Geociências, Engenharia Biomédica - Bacharelado, 2024.
Inclui referências.

1. Aprendizado de Máquina. 2. Biomarcadores moleculares. 3. Câncer. I. Santos, Wellington Pinheiro dos . (Orientação). II. Santana, Maíra Araújo de . (Coorientação). IV. Título.

620 CDD (22.ed.)

**UNIVERSIDADE FEDERAL DE PERNAMBUCO**

**CENTRO DE TECNOLOGIA E GEOCIÊNCIAS**

**DEPARTAMENTO DE ENGENHARIA BIOMÉDICA**

WHENDEL MUNIZ DOS SANTOS

# CLASSIFICAÇÃO DE TIPOS DE CÂNCER EM MULHERES A PARTIR DE APRENDIZADO DE MÁQUINA E SEQUENCIAMENTOS GENÉTICOS

Trabalho Supervisionado, apresentado ao Curso de Engenharia Biomédica do Centro de Tecnologia e Geociências da Universidade Federal de Pernambuco, como requisito para a obtenção do título de Bacharel em Engenharia Biomédica.

Aprovado em: 17/ 07/ 2024 .

**BANCA EXAMINADORA**

_____

Prof. Dr. Wellington Pinheiro dos Santos (Orientador)
Universidade Federal de Pernambuco

_____

Profa. Dra. Maíra Araújo de Santana (Coorientadora)
Universidade Federal de Pernambuco

_____

Prof. Dr. Fabiano Tonaco Borges
Universidade Federal de Mato Grosso

_____

Prof. Dr. Fábio Henrique Cavalcanti de Oliveira
Universidade de Pernambuco

## RESUMO

O câncer é considerado uma das doenças que mais causam mortes no mundo e, especificamente, assola a população feminina no Brasil, como nos casos do Câncer de Mama. Nesse contexto, o país apresenta problemas de infraestrutura com relação ao diagnóstico precoce da doença, afetando várias mulheres, principalmente aquelas que estão em vulnerabilidade social e econômica. Na contemporaneidade, abordagens utilizando Inteligência Artificial e sequenciamentos genéticos são estudadas para predizer e classificar tipos de câncer, permitindo a celeridade de detecção. Dessa forma, esse trabalho pretende, por meio de aprendizado de máquina e biomarcadores moleculares, classificar tipos de câncer que são mais incidentes em mulheres brasileiras nos últimos anos. Os dados de quantificação de genes em tecidos cancerosos, provenientes da *Broad Institute*, localizada em *Cambridge, Massachusett*, foram comparados com os gráficos de incidência de câncer em mulheres, disponibilizado pelo Instituto Nacional de Câncer (INCA), a fim de selecionar uma amostra para o trabalho. Logo após, a amostra foi dividida em 6 subamostras, baseadas na relevância dos genes de referência, pela validação no algoritmo *Random Forest*. A partir disso, as subamostras foram divididas em bases de treinamento e de teste para implementação de 5 classificadores: *Random Forest*, SVM, J48, *Bayes Network* e *Naive Bayes*. Por fim, foram analisados os desempenhos dos algoritmos de classificação, além de testes de validação com mais interações para discutir a relevância de genes comuns para diagnóstico do câncer de mama. Espera-se contribuir como potencial auxílio na detecção de tipos de câncer em mulheres. Além disso, esse trabalho pretende incentivar a pesquisa, por engenheiros biomédicos, sobre a importância de biomarcadores moleculares para identificação de doenças.



**Palavras-chave:** Aprendizado de Máquina; Biomarcadores moleculares; Câncer.

## ABSTRACT

Cancer is considered one of the diseases that cause the most deaths in the world and, specifically, ravages the female population in Brazil, such as breast cancer. In this context, the country presents infrastructure problems in relation to early diagnosis of the disease, affecting several women, especially those who are socially and economically vulnerable. Nowadays, approaches using Artificial Intelligence and genetic sequencing are studied to predict and classify types of cancer, allowing for faster detection. Therefore, this work intends, through machine learning and molecular biomarkers, to classify types of cancer that are most common in Brazilian women in recent years. Gene quantification data in cancerous tissues, from the Broad Institute, located in Cambridge, Massachusett, were compared with cancer incidence graphs in women, made available by the National Cancer Institute (INCA), in order to select a sample for the work. Soon after, the sample was divided into 6 subsamples, based on the relevance of the reference genes, by validation in the Random Forest algorithm. From this, the subsamples were divided into training and testing bases to implement 5 classifiers: *Random Forest*, SVM, J48, *Bayes Network* and *Naive Bayes*. Finally, the performance of the classification algorithms was analyzed, in addition to validation tests with more interactions to discuss the relevance of common genes for the diagnosis of breast cancer. It is expected to contribute as a potential aid in the detection of types of cancer in women. Furthermore, this work aims to encourage research, by biomedical engineers, on the importance of molecular biomarkers for identifying diseases.



**Keywords:** Machine Learning; Molecular Biomarkers; Cancer.

## LISTA DE FIGURAS

## LISTA DE GRÁFICOS

## LISTA DE TABELAS

## LISTA DE ABREVIAÇÕES

| | |
|---|---|
| AUC-ROC | Área sob a curva ROC |
| BRCA | Carcinoma invasivo de mama |
| CNN | Redes Neurais de Convolução |
| COAD | Adenocarcinoma de cólon |
| ctDNA | DNA circulante de tumor |
| DNA | Ácido desoxirribonucleico |
| INCA | Instituto Nacional de Câncer |
| KNN | *K-nearest neighbor* |
| LR | Regressão Logística |
| LUAD | Adenocarcinoma pulmonar |
| MAPK/ERK | *Mitogen-Activated Protein Kinase/Extracellular Signal-Regulated Kinase* |
| MET | Fator de Transição Mesenquimal-epitelial |
| *Non*-TNBC | Não Triplo-Negativo |
| NGB | *Naive Bayes* |
| PCA | Análise de Componentes Principais |
| RAS | *Rat Sarcoma* |
| READ | Adenocarcinoma de Reto |
| RF | *Random Forest* |
| RNA | Ácido ribonucleico |
| RET | *Rearranged during Transfection* |
| rL-GenSVM | Máquina de Vetor de Suporte Linear Multi-class |

| | |
|---|---|
| SUS | Sistema Único de Saúde |
| TCGA | *The Cancer Genome Atlas* |
| THCA | Carcinoma de Tireoide |
| TNBC | Triplo-Negativo |
| VEGFC | Fator C de Crescimento Endotelial Vascular |

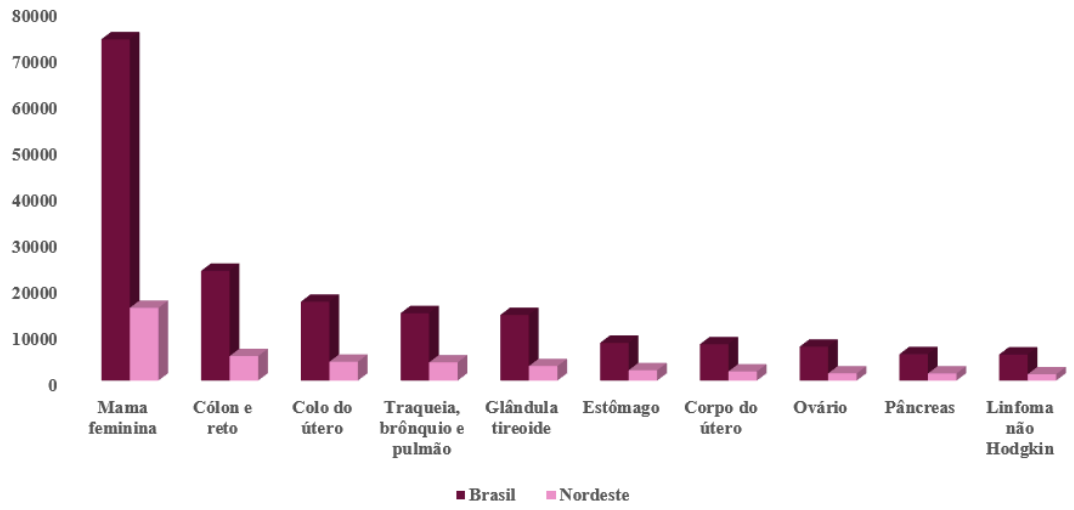
**SUMÁRIO**

# 1 INTRODUÇÃO

## 1.1 MOTIVAÇÃO

O câncer é considerado uma das doenças que mais causam mortes no mundo e, especificamente nas Américas, é esperado que o índice de mortalidade cresça até 2,1 milhões até 2030 (Organização Pan-Americana da Saúde, 2020). Em 2022, o câncer de mama feminino foi o segundo mais incidente do mundo com 2,3 milhões de casos, cerca de 11,6% das notificações totais (Organização Pan-Americana da Saúde, 2024). Nesse contexto, os casos de câncer em mulheres no Brasil é um tema bastante pertinente e de atenção para a saúde coletiva. Em 2023, a estimativa de notificações de casos de cânceres na população feminina foi de aproximadamente 362,7 mil (Instituto Nacional do Câncer, 2023). O câncer de mama foi o mais incidente, alcançando 30,1% dos casos totais em mulheres, seguido pelo câncer de cólon e reto, e de colo de útero como os 3 mais notificados (Instituto Nacional do Câncer, 2023). Ademais, a estimativa de distribuição dos 10 tipos de câncer mais incidentes em mulheres, no ano de 2023, no Brasil e na região Nordeste pode ser visualizada na Figura 1.

Figura 1 – Estimativa de incidência de câncer em mulheres no ano de 2023




Fonte: Instituto Nacional de Câncer (2023)

Nota: Figura elaborada com base nos dados informados pelo Instituto Nacional de Câncer.

Portanto, a partir da apresentação estatística desse cenário, há a necessidade de reforçar o sistema de saúde brasileiro com tecnologias que prezem pelo diagnóstico precoce

de cânceres em mulheres para propiciar um tratamento mais precoce, certamente menos traumático e menos custoso às pacientes e ao Estado. Ademais, a utilização do Aprendizado de Máquina (AM) para classificação de tipos de câncer por meio de sequenciamentos genéticos já é bastante vista na literatura e pode ser uma alternativa para auxílio na área da oncologia (Wang et al., 2021).

## 1.2 JUSTIFICATIVA

No Brasil, em 2021, cerca de 62% dos casos de câncer foram diagnosticados em estágios avançados da doença (Instituto Oncoguia, 2023). Dessa forma, é necessária a preocupação com o diagnóstico precoce de câncer no Brasil, visto que muitas técnicas implementadas, atualmente, ainda apresentam gargalos quanto ao tempo de descoberta da doença e expectativa de cura após o diagnóstico (Instituto Oncoguia, 2023). Ademais, para adiantar o diagnóstico e com uma abordagem menos invasiva, a detecção de genes biomarcadores e até proteínas encontradas em células tumorais são alternativas para auxílio desse monitoramento oncológico. Além da falta de tecnologias suficientes para detecção do câncer atualmente, o impacto disso nas mulheres pobres e moradoras de periferia é um sinal de urgência. Pesquisas apontam que 7 em cada 10 moradores de favelas não tem acesso a diagnósticos e exames (Instituto Oncoguia, 2023). Nesse contexto, o uso do AM para auxiliar mulheres que apresentem condições mais precárias e de alta vulnerabilidade social no diagnóstico de câncer é extremamente necessário, principalmente no Sistema Único de Saúde (SUS).

## 1.3 OBJETIVO

Este trabalho tem como objetivo geral construir modelos de aprendizado de máquina para detecção precoce dos tipos de câncer mais comuns em mulheres a partir de sua representação na forma da quantificação de biomarcadores moleculares em amostras.

### 1.3.1 Objetivos Específicos

1. Selecionar a base de dados utilizada para os experimentos de AM a partir dos dados estatísticos do Instituto Nacional de Câncer (INCA) sobre incidência de câncer em mulheres;

2. Analisar a relevância dos biomarcadores e separá-los em subamostras;
3. Separar cada subamostra em conjuntos de treinamento e teste;
4. Analisar o desempenho do conjunto de treino nos algoritmos pré-estabelecidos e extrair o melhor entre eles para o conjunto de teste;
5. Analisar o desempenho do conjunto de teste com o algoritmo treinado e escolhido;
6. Avaliar a qualidade dos resultados obtidos e a relevância para o cenário atual de diagnóstico de câncer em mulheres.

## 1.4 ORGANIZAÇÃO DO TRABALHO

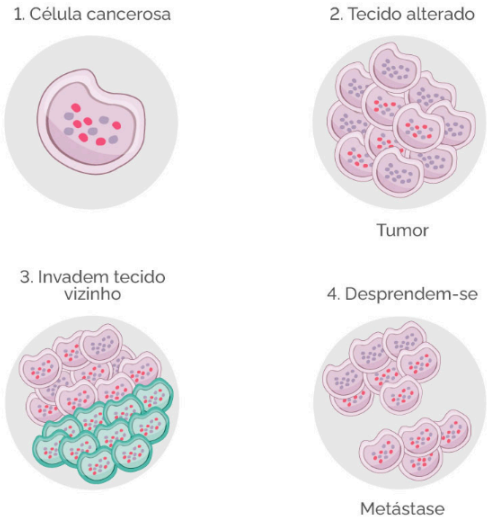
O presente trabalho está organizado na seguinte estrutura: no capítulo 2, será apresentada a fundamentação teórica para basear as principais definições no estudo do câncer, como o entendimento de tumores, da genômica do câncer e exemplos da literatura de biomarcadores moleculares para diagnóstico de alguns tipos de câncer. Ademais, também serão exemplificados artigos e projetos de pesquisa que utilizaram inteligência artificial para prever ou classificar essas doenças, além de definições de aprendizado de máquina e indicadores estatísticos utilizados nas etapas seguintes. Logo após, será apresentada a metodologia utilizada desde a aquisição dos dados até as fases de treinamento e teste com diferentes algoritmos de classificação no capítulo 3. Por fim, nos capítulos 4 e 5 serão apresentados os resultados relevantes desse trabalho, além de discussões sobre o desempenho dos classificadores utilizados e possível utilizado desse estudo para a saúde coletiva voltada para as mulheres brasileiras.

# 2 FUNDAMENTAÇÃO TEÓRICA

## 2.1 CÂNCER

O termo câncer é utilizado, na oncologia, para definir as doenças causadas pelo crescimento desordenado das células em um determinado tecido do corpo (Instituto Nacional de Câncer, 2022). A partir de sucessivas divisões celulares, formam-se tumores que podem invadir órgãos vizinhos à origem do câncer e, também, regiões mais distantes por meio da circulação sanguínea e linfática, sendo esta última a fase mais severa e denominada de metástase (Patel, 2020). A Figura 2 ilustra a progressão do câncer no corpo humano de maneira geral.

Figura 2 – Representação de progressão de células cancerígenas em órgãos




Fonte: Instituto Nacional de Câncer (2022)

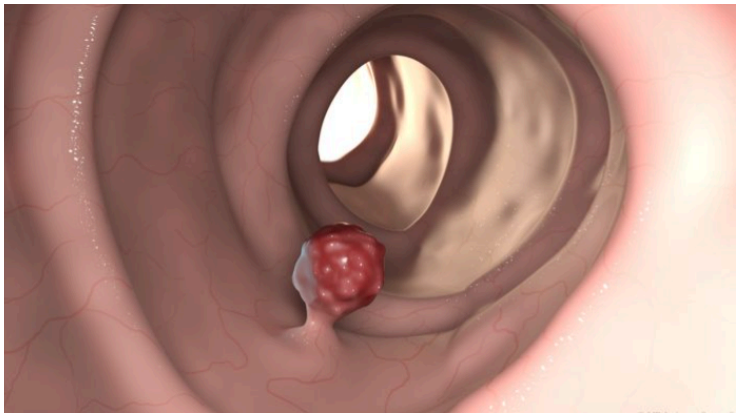
### 2.1.1 Tumores

Os tumores podem ser diferenciados em benignos e malignos. O tumor benigno apresenta limites de crescimento bem definidos e células que se assemelham com a de origem, não havendo riscos de atingir tecidos vizinhos ou de desenvolver metástases (Patel, 2020). Entre os exemplos mais comuns desse caso, há os adenomas originários em glândulas corpóreas mais generalizadas, os miomas encontrados no tecido muscular liso do útero e os

lipomas que são camadas de gordura encontradas entre a pele e os tecidos musculares do tronco, cabeça e membros adjacentes (Hospital Israelita Albert Einstein, 2023).

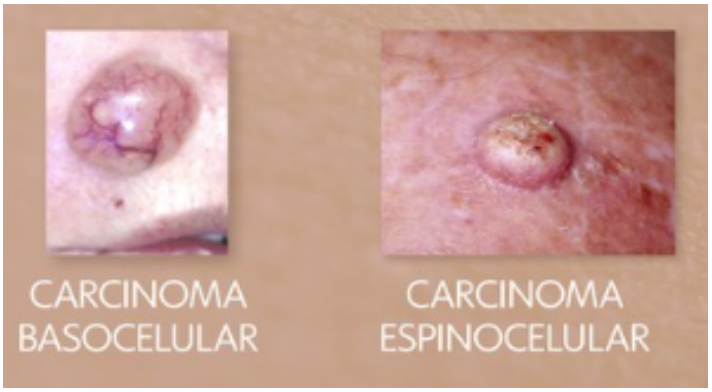
Por outro lado, os tumores malignos não apresentam uma expectativa de crescimento delimitada, além da falta de semelhança entre as células atuantes e a originária (Patel, 2020). A partir disso, essas neoplasias podem desenvolver diferentes tipos de câncer, além de metástases irreversíveis a depender do período de diagnóstico (Patel, 2020). Dentre as categorias principais de tumor maligno, pode-se citar o carcinoma que é caracterizado pela origem de crescimento na pele e em células de glândulas, sendo este último chamado de adenocarcinoma (Pfizer Brasil, 2022). Ademais, há o sarcoma, o qual surge em células do tecido conjuntivo como músculos, ossos e cartilagens, além da leucemia, conhecida por invadir células do tecido sanguíneo, podendo ser diferenciada em mielomas e linfomas (Pfizer Brasil, 2022). As Figuras 3 e 4 ilustram o aparecimento de adenocarcinomas e carcinomas em órgãos humanos.

Figura 3 – Representação ilustrativa do adenocarcinoma em cavidades



Fonte: Demarco (2021)

Figura 4 – Exemplo de carcinomas na pele

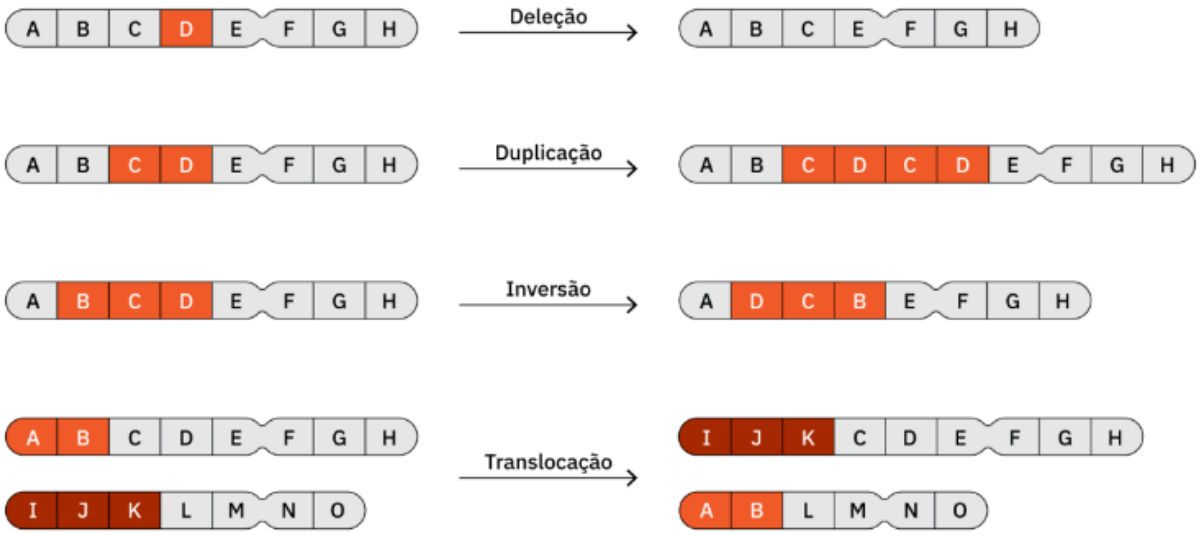



Fonte: G1 (2019)

### 2.1.2 Genômica do Câncer

O surgimento e desenvolvimento do câncer nos seres humanos também pode ser explicado à nível molecular. O ácido desoxirribonucleico (DNA) é uma estrutura, em formato de dupla hélice, formada por uma sequência não-aleatória de milhares de nucleotídeos e pareada por 4 bases nitrogenadas: adenina, citosina, timina e guanina (Alberts et al., 2017). Esse sequenciamento carrega a informação genética de uma célula, também denominada de gene, que será transcrita em ácido ribonucleico (RNA) e traduzida em proteínas, finalizando o processo de expressão gênica (Alberts et al., 2017). A partir desse contexto, quando há alterações no procedimento de duplicação do DNA, afetando o pareamento de bases, arranjo de nucleotídeos no sequenciamento genético ou outros processos biológicos, alguns genes sofrem mutações que podem modificar células e, a partir de múltiplas divisões, transformá-las em tumores (Instituto Nacional de Câncer, 2022). Dentre as alterações genéticas no DNA mais comuns na literatura, podem ser mencionadas as mutações de deleção, duplicação, inversão e translocação, como podem ser ilustradas na Figura 5.

Figura 5 – Alterações Cromossômicas Estruturais




Fonte: Faria (2021)

Os genes modificados que podem desencadear o surgimento do câncer são classificados em 3 grupos: proto-oncogenes, genes supressores de tumor e genes de reparo de DNA. Os proto-oncogenes regulam a divisão celular, porém quando são mutados, se tornam oncogenes e todas as células que são comandadas por eles recebem a instrução de se dividirem de maneira desordenada e ininterrupta, contribuindo para o desenvolvimento do câncer no tecido ou órgão de origem (Instituto Nacional de Câncer, 2022). Já os genes

supressores de tumor são responsáveis por inibir esse processo de divisão das células, mas se forem desregulados, podem perder a comunicação com elas e, consequentemente, auxiliar no surgimento de tumores (Kontomanolis et al., 2020; Otmani; Lewalle, 2021). Os genes de reparo de DNA inspecionam possíveis erros de pareamento de bases nitrogenadas e processos adjuntos, porém com alterações causadas por mutação, os erros podem se acumular, também gerando o aparecimento do câncer (Moubarz et al, 2022). Essas informações genéticas modificadas são consideradas marcadores biológicos, ou seja, genes que podem facilitar a diferenciação entre células cancerosas e saudáveis (Khan et al., 2022).

#### *2.1.2.1 Uso de Marcadores Biológicos na Identificação de Cânceres na Literatura*

No fim da década de 1970, moléculas específicas encontradas em fluidos corporais, como sangue e urina, começaram a ser validadas como biomarcadores para diagnósticos de doenças (Khan et al., 2022). Com o avanço dos estudos oncológicos, o uso de genes para identificação de cânceres se tornou cada vez mais presente na genômica, tendo um bom primeiro exemplo a identificação do gene BRAF em DNA circulante de tumor (ctDNA), um fragmento de DNA liberado por células cancerosas na corrente sanguínea (Oliveira et al., 2022). O BRAF codifica a proteína B-Raf, que participa da via de sinalização *Mitogen-Activated Protein Kinase/Extracellular Signal-Regulated Kinase* (MAPK/ERK), responsável pela regulação de crescimento celular e apoptose (morte celular programada) (Halle; Johnson, 2021). A partir da mutação BRAF V600E, que é caracterizada pela troca do aminoácido valina por ácido glutâmico na posição 600 da sequência genética, a ativação de crescimento celular será contínua, corroborando para o surgimento de possíveis tumores (Halle; Johnson, 2021). A partir desse contexto, Oliveira et al. apontam que a expressão desse gene no ctDNA pode ser útil para o monitoramento da resposta do câncer de pele ao tratamento por imunoterapia (Oliveira et al, 2022).

Nos estudos de Câncer de Mama, as mutações dos genes BRCA1 e BRCA2, associados a falhas em genes de reparo de DNA, são potenciais identificadores e preditores da doença, tendo prevalência de 22,73% em pacientes com câncer de mama com histórico familiar dessa doença (Han; Kim, 2021). Esses genes são responsáveis por produzir proteínas que auxiliam na manutenção de DNAs danificados, mantendo a estabilidade genética das células (Kim; Shin, 2021). A partir de alterações no BRCA1 e BRCA2, o reparo é desfalcado e o acúmulo de erros torna a célula suscetível a se tornar um tumor maligno. Ademais, outro

gene de reparo genômico importante nesse contexto é o *Nibrin* (NBN), o qual tem como função principal transportar as proteínas MRE11A e RAD50 até o local de quebra da dupla hélice do DNA com o objetivo de reconstruí-la (Nithya; Chandrasekar, 2019). A partir de mutações no NBN, há a teoria de que os erros provenientes da replicação também possam não ser percebidos corretamente, contribuindo para o descontrole de crescimento celular. A partir desse contexto, Nithya e Chandrasekar analisaram mudanças em aminoácidos em regiões estratégicas do *Nibrin*, como por exemplo, a troca de fenilalanina em leucina na posição 266 do gene, e perceberam que, neste caso, há o aparecimento de anemia plástica e síndrome de quebra de *Ninjmegan*, caracterizada como um tipo de microcefalia (Nithya; Chandrasekar, 2019).

Nos estudos do Câncer Colorretal, também é possível encontrar biomarcadores em potencial para identificação da doença, como os genes da família do Fator C de Crescimento Endotelial Vascular (VEGFC) (Malki et al., 2020). Os VEGFCs são responsáveis por promover o crescimento de novos sanguíneos, também chamado de angiogênese. Esse fenômeno biológico é essencialmente vinculado ao crescimento de tumores e até de metástase, correlacionando com os fatores de crescimento vascular endotelial (Malki et al., 2020). Ademais, os genes da família VEGF são bastante expressados em fases mais avançados do Câncer Colorretal, mas também podem ser identificados na fase de adenoma (Malki et. al, 2020). No caso do Câncer de Pulmão, genes como o Fator de Transição Mesenquimal-epitelial (MET), que tem como função codificar o receptor tirosina-quinase para auxiliar no crescimento e diferenciação celular, e o gene BRAF supracitado são bastante utilizados na oncologia para identificar níveis da doença (Sears; Mazzone, 2020). Por fim, o Câncer de Tireóide também pode ser identificado e analisado pelo gene codificante da proteína B-Raf, além de outros genes como o *Rat Sarcoma* (RAS), que regula o crescimento celular, e também o proto-oncogene *Rearranged during Transfection* (RET), agindo de maneira similar ao RAS (Abdullah et al., 2019).

## 2.2 O USO DA INTELIGÊNCIA ARTIFICIAL PARA CLASSIFICAÇÃO DE CÂNCER A PARTIR DE SEQUENCIAMENTOS GENÉTICOS – UMA BREVE REVISÃO BIBLIOGRÁFICA

Na atualidade, diversos trabalhos científicos estão contribuindo para as descobertas do uso da inteligência artificial para predição e classificação de câncer por sequenciamentos genéticos. Um bom exemplo introdutório é o uso de aprendizado profundo para classificação de câncer a partir de conformação bidimensional (2D) dos genes de referência em expressão das doenças (Wang et al., 2021). Nessa pesquisa, Wang et al. desenvolveram um método de classificação de cânceres baseado em redes neurais de convolução (CNN), nomeado de *MI_DenseNetCAM*, com amostras de 33 tipos de câncer, como o câncer de Mama, Colorretal, de Pulmão e, também, de Tireoide. As amostras foram coletadas de modo experimental e os genes de referência foram extraídos por meio de um processo chamado *RNA-Seq* para depois serem quantificados (Wang et al., 2021). Logo após, os dados foram pré-processados para padronização, transformados em uma matriz 2D e, por fim, foram colocados no modelo de classificação (Wang et al., 2021). O modelo de convolução apresenta camadas densas entre as redes neurais para evitar *overfitting* para menores instâncias, além de utilizar o Grad-cam para, a partir dos mapas de calor gerados a cada experimento, identificar quais genes foram os mais relevantes para a classificação (Wang et al., 2021). Como resultados, o modelo proposto por Wang et al. obteve acurácia de 96,81% em comparação com o algoritmo de máquina de vetor de suporte linear multi-class (rL-GenSVM) com 87,29%.

Outra abordagem significativa encontrada na literatura foi no uso de CNNs para predição de genes importantes para determinar câncer de pâncreas especificamente (Mori et al., 2021). A partir dessa hipótese, Mori et al. coletaram amostras de tecidos com o adenocarcinoma pancreático e saudáveis para extraírem os genes e utilizar em uma rede neural artificial. Diferentemente de Wang et al., neste trabalho os dados não foram transformados em uma matriz 2D, e sim foram configurados como uma planilha de dados (Mori et al., 2021). Os pesquisadores utilizaram uma CNN adaptada com uma camada de seleção de atributos antes das camadas de convolução. Essa camada foi responsável por dar diferentes pesos para as entradas, que seriam os genes, tendo uma classificação dos mais importantes ao final de cada treinamento (Mori et al., 2021). Mori et al. não utilizaram métodos estatísticos como forma de validar a classificação dos genes, e sim a contagem daqueles que mais estiveram na posição primordial de relevância para o treino do modelo (Mori et al., 2021). A partir desse método, eles encontraram genes relevantes que,

comparados com a literatura, eram bastante estudados na oncologia para o câncer pancreático (Mori et al., 2021).

Com relação ao Câncer de Mama, trabalhos de pesquisa relacionados à classificação de subtipos da doença podem ser encontrados na literatura, como o uso de Aprendizado de Máquina para diferenciar o Triplo-Negativo (TNBC) do Não Triplo-Negativo (*Non*-TNBC) (Wu; Hicks, 2021). Wu e Hicks coletaram dados de *RNA-Seq* contendo 110 amostras de câncer do tipo TNBC e 992 do tipo *Non*-TNBC a partir do *The Cancer Genome Atlas* (TCGA) e logo após as instâncias foram normalizadas e finalizadas em 116 e 818 para TNBC e *Non*-TNBC respectivamente. A partir dos dados já pré-processados, os autores utilizaram quatro modelos de AM: SVM, *K-nearest neighbor* (KNN), *Naive Bayes* (NGB) e Árvores de Decisão (DT) para escolha do melhor algoritmo e estudo do treinamento e teste da base de dados posteriormente (Wu; Hicks, 2021). Os resultados encontrados foram interpretados estatisticamente e apresentados, em termos de acurácia, a classificação dos modelos foram de 87% para KNN, 85% para NGB, 87% para DT e 90% para SVM (Wu; Hicks, 2021). A partir dessa análise prévia, Wu e Hicks utilizaram o SVM como algoritmo para os experimentos de treinamento e teste, chegando a 90% de acurácia para os agrupamentos de treino e 82% para os agrupamentos de teste.

Outra abordagem importante nos estudos do Câncer de Mama na literatura é o de classificação de Carcinomas Ductais Invasivos por meio de AM (Roy et al., 2020). Roy et al. obtiveram os dados a partir de dados de *RNA-Seq* a partir do TCGA e dividiram as amostras em 2 grupos: Estágio Inicial e Estágio Final, contendo 2 subgrupos cada (Estágio I e II para o grupo de Estágio Inicial e Estágio III e IV para o grupo de Estágio Final). No total, 610 amostras foram coletadas, sendo 107 do Estágio I, 362 do Estágio II, 128 do Estágio III e 13 do Estágio IV (Roy et al., 2020). Os dados foram separados em conjunto de treino e teste por 80% e 20% amostras respectivamente, utilizando o *Random Forest* (RF), DT, NB, Regressão Logística (LR) e SVM. Os autores observaram que o RF obteve a melhor peformance em todos os agrupamentos de treino, tendo 95% de acurácia para validação (Roy et al., 2020).

## 2.3 APRENDIZADO DE MÁQUINA

O aprendizado de máquina é considerado um ramo advindo da Inteligência Artificial, que tem como objetivo colaborar com a inteligência humana para melhorar análises de classificação e predição de dados por meio de algoritmos computacionais (Helm et al., 2020). Esses algoritmos, a partir de uma quantidade grande de dados como entradas e saídas, são

configurados para aprender a reconhecer padrões e, dessa forma, classificar e até predizer, em alguns casos, outros dados considerados desconhecidos pelo sistema criado (Helm et al., 2020). Ademais, a partir de repetições significativas de treinamento desse algoritmo configurado, a máquina será capaz de receber um conjunto de dados de entrada e gerar uma saída baseada em uma decisão para resolver um problema específico (Helm et al., 2020). O desempenho da máquina é analisado a partir de métricas diversas, sendo a estatística a mais comum na literatura, e na seção 2.3.4 as métricas utilizadas nesse trabalho serão descritas de maneira mais aprofundada.

O aprendizado de máquina pode ser dividido em 3 subáreas: aprendizado não supervisionado, aprendizado supervisionado e aprendizado por reforço. O aprendizado não supervisionado é uma técnica que carece de uma variável de saída, sendo uma forma de encontrar relações entre a base de dados sem necessariamente ter um resultado medido (Jiang; Gradus; Rosellini, 2020). Como exemplos, esse tipo de método de aprendizado já foi citado na literatura para uso em transtornos mentais, análise de componentes principais (PCA) e análise fatorial (Jiang; Gradus; Rosellini, 2020).

Em contrapartida, o aprendizado supervisionado tem como objetivo predizer e classificar um resultado específico que se espera pelos administradores da máquina (Jiang; Gradus; Rosellini, 2020). A gama de uso desse tipo de aprendizado na literatura é bastante significativa, como em prognóstico de doenças, como o próprio câncer, sendo base de dados com um único câncer ou até com vários (Jiang; Gradus; Rosellini, 2020). A aplicação de classificação e predição se diferencia pela conformação dos dados: se os dados estiverem separados por categorias, a classificação é a mais adequada, mas se os dados estiverem agrupados de forma contínua, a predição é a mais utilizada (Freedman, 2009). Por fim, o aprendizado por reforço é uma técnica que utiliza aspectos das dos dois tipos de aprendizado supracitados, treinando a máquina para atingir melhores resultados por meio do método de tentativa e erro (Amazon Web Services, 2023). A partir disso, o algoritmo descobre, por meio de feedback próprio, quais os melhores caminhos de decisão utilizando o paradigma de gratificação, ou seja, um valor positivo, negativo ou nulo para cada ação feita ambientada por entradas e saídas de dados (Amazon Web Services, 2023).

### 2.3.1 Métricas Estatísticas de Desempenho

As métricas Estatísticas de Desempenho utilizadas nesse trabalho são baseadas em variáveis utilizadas na decisão e classificação obtida pelos algoritmos. Considerando *t* uma classe descrita na base de dados, essas variáveis podem ser descritas da seguinte forma:

1. Verdadeiros Positivos (VP□): classificações corretas obtidas pelo algoritmo, ou seja, dados classificados como classe *t* e que realmente pertencem à classe *t*;
2. Falsos Positivos (FP□): classificações incorretas obtidas pelo algoritmo, ou seja, dados classificados como classe *t*, mas que não pertencem à classe *t*;
3. Verdadeiros Negativos (VN□): classificações corretas obtidas pelo algoritmo, porém neste caso os dados não são classificados como classe *t* e realmente não pertencem à classe *t*;
4. Falsos Negativos (FN□): classificações incorretas obtidas pelo algoritmo, porém nesse caso os dados não são classificados como classe *t*, mas pertencem à classe *t*;
5. Total de classificações (P).

#### *2.3.1.1 Acurácia*

A acurácia mede a taxa de proximidade da classificação dos dados em comparação à classificação real deles. A fórmula para esse indicador pode ser descrita por (2.1):

$$\text{Acurácia} = \frac{\sum_{t=1}^{n} VP_t}{P} \tag{2.1}$$

sendo $n$ o número de classes.

#### *2.3.1.2 Índice Kappa*

O Índice Kappa (K) mede a concordância entre classificadores que categorizam um conjunto de itens em um determinado número de classes. O medidor K pode ter os seguintes valores de intervalos críticos:

1. K = 1: concordância ideal;
2. K = 0: concordância esperada para o acaso;

3. K < 0: discordância (concordância menor que a esperada).

Os valores de K podem estar entre -1 e 1. Logo, quanto mais próximo de 1, mais ideal a concordância (Mchugh, 2012). Esse índice pode ser calculado a partir da matriz confusão n x n das probabilidades observadas e esperadas, além de $C_{tu}$ que é a quantidade de vezes que a classe t foi classificada como classe u. A partir disso, a expressão matemática para esse medidor pode ser vista em (2.2):

$$K = \frac{P_{o_{-P_e}}}{1-P_e} \tag{2.2}$$

em que $P_o$ é a probabilidade observada calculada por (2.2.1):

$$P_o = \frac{\sum_{t=1}^{n} C_{tt}}{P} \tag{2.2.1}$$

e $P_e$ é a probabilidade esperada calculada por (2.2.2):

$$P_e = \frac{\sum_{i=1}^{n}((\sum_{u=1}^{n} C_{tu}) \, x \, (\sum_{u=1}^{n} C_{ut})}{P^2} \tag{2.2.2}$$

em que n é o número de classes, $C_{tt}$ é o número de vezes em que t classes foram classificadas corretamente, $C_{tu}$ o total esperado da classe t e $C_{ut}$ o total observado para a classe t.

*2.3.1.3 Sensibilidade*

A sensibilidade mede a proporção dos verdadeiros positivos com relação a todas as outras classificações obtidas na classe de positivos (verdadeiros e falsos). A fórmula para esse medidor pode ser vista em (2.3):

$$\text{Sensibilidade} = \frac{1}{n}\sum_{t=1}^{n}\frac{VP_t}{VP_t+FN_t} \tag{2.3}$$

sendo $n$ o número de classes.

*2.3.1.4 Especificidade*

A especificidade mede a proporção dos verdadeiros negativos com relação a todas as classificações obtidas na classe de negativos (verdadeiros e falsos). A fórmula para esse medidor pode ser vista em (2.4):

$$\text{Especificidade} = \frac{1}{n}\sum_{t=1}^{n}\frac{VN_t}{VN_t+FP_t} \tag{2.4}$$

sendo $n$ o número de classes.

*2.3.1.5 Área sob a curva ROC*

Á área sob a curva ROC (AUC-ROC) mede o desempenho do algoritmo em distinguir as classes positivo e negativo (Moorthy; Sankar, 2019). O AUC-ROC pode estar entre os seguintes intervalos numéricos:

1. AUC-ROC = 1: um modelo ideal que classifica todas as instâncias de maneira correta;
2. AUC-ROC = 0.5: um modelo que classifica as instâncias da mesma forma que uma classificação aleatória;
3. AUC-ROC < 0.5: um modelo que classifica as instâncias de maneira inferior à classificação aleatória.

Essa medida pode ser calculada a partir de (2.5):

$$\text{AUC} = \sum_{t=1}^{n}\left(\frac{VPR_t - VPR_{t+1}}{2}\right)x\,(FPR_t - FPR_{t+1}) \tag{2.5}$$

em que $TPR$ é a sensibilidade, $FPR$ a especificidade e $n$ o número de classes.

# 3 METODOLOGIA

## 3.1 A BASE DE DADOS

Os dados para esse trabalho foram extraídos diretamente da *Broad Institute*, localizada em *Cambridge, Massachusetts*. A base de dados possui 10446 amostras de tecidos categorizados por 33 tipos de câncer, que são quantificados por 20531 genes de referência pelo procedimento *RNASeq* sem normalização estatística. A Tabela 1 apresenta as siglas referentes a cada câncer por localização tecidual, além da quantidade de amostras para cada doença.

Tabela 1 – Catalogação e distribuição das amostras de câncer

| CONFIGURAÇÃO | DESCRIÇÃO | QUANTIDADE DE AMOSTRAS |
|---|---|---|
| ACC | Carcinoma adrenocortical | 79 |
| BLCA | Carcinoma Urotelial da Bexiga | 427 |
| BRCA | Carcinoma invasivo da mama | 1212 |
| CESC | Cânceres cervicais e endocervicais | 309 |
| CHOL | Colangiocarcinoma | 45 |
| COAD | Adenocarcinoma de cólon | 328 |
| DLBC | Neoplasia Linfóide Linfoma Difuso de Grandes Células B | 48 |
| ESCA | Carcinoma de esôfago | 196 |
| GBM | Glioblastoma multiforme | 171 |
| HNSC | Carcinoma espinocelular de cabeça e pescoço | 566 |
| KICH | Cromófobo renal | 91 |
| KIRC | Carcinoma renal de células claras renal | 606 |
| KIRP | Carcinoma de células papilares renais renal | 323 |
| LAML | Leucemia mielóide aguda | 173 |
| LGG | Glioma cerebral de grau inferior | 530 |
| LIHC | Carcinoma hepatocelular do fígado | 423 |
| LUAD | Adenocarcinoma pulmonar | 576 |
| LUSC | Carcinoma de células escamosas do pulmão | 552 |
| MESO | Mesotelioma | 87 |
| OV | Cistadenocarcinoma seroso ovariano | 307 |
| PAAD | Adenocarcinoma Pancreático | 183 |
| PCPG | Pheochromocytoma and Paraganglioma | 187 |
| PRAD | Adenocarcinoma pancreático | 550 |
| READ | Adenocarcinoma de reto | 105 |
| SARC | Sarcoma | 265 |
| SKCM | Melanoma cutâneo da pele | 473 |
| STAD | Adenocarcinoma de estômago | 450 |
| TGCT | Tumores de células germinativas testiculares | 156 |
| THCA | Carcinoma de tireóide | 568 |
| THYM | Timoma | 122 |
| UCEC | Carcinoma Endometrial do Corpo Uterino | 201 |
| UCS | Carcinossarcoma Uterino | 57 |
| UVM | Melanoma Uveal | 80 |

Fonte: Broad GDAC Firehose (2016)

Nota: Tabela elaborada com base nos dados informados pelo *Broad GDAC Firehose*.

## 3.2 REARRANJO DA BASE DE DADOS

Primeiramente, a base de dados foi extraída separadamente por cada tipo de câncer e em formato *data*. Logo após, os dados foram transformados em uma planilha do tipo *csv*, concatenados e transpostos para que as instâncias (tecidos com câncer) estivessem em linhas e os atributos (genes quantificados) em colunas. Por fim, foi adicionada uma coluna para identificar cada amostra pelo respectivo tipo de câncer.

## 3.3 AMOSTRA DE TIPOS DE CÂNCER EM MULHERES

A partir da base de dados completa e do gráfico de incidência de câncer em mulheres no Brasil, disponibilizado pelo Instituto Nacional de Câncer (Figura 1), foram escolhidos os 5 tipos de câncer mais comuns na população feminina como amostra para as próximas etapas do projeto.

## 3.4 ANÁLISE ESTATÍSTICA DESCRITIVA DA AMOSTRA

A amostra foi submetida a uma análise estatística descritiva para verificar a distribuição dos dados e possíveis padrões de agrupamentos. Essa etapa foi feita a partir de programação em *Python* pelo *Google Colab*.

## 3.5 SUBAMOSTRAGENS DA BASE DE DADOS DE TRABALHO

A base de dados escolhida foi submetida a um treinamento de validação com o classificador *Random Forest* (n = 10), pelo software *Weka*, para observar a porcentagem de relevância dos atributos. A partir disso, foram feitas 6 subamostras baseadas na relevância dos genes: acima de 0%, acima de 10%, acima de 20%, acima de 30%, acima de 40% e acima de 50%.

## 3.6 PRÉ-TREINAMENTO DAS SUBAMOSTRAGENS

Após a divisão das subamostras, foram feitas as separações da base de treinamento e de teste, 70% e 30% respectivamente, pelo filtro supervisionado *Resample* no software *Weka*.

Ademais, utilizando o filtro também supervisionado *SMOTE*, incluído no programa supracitado, foi feito o balanceamento das 6 bases de treino. O percentual de aumento de um conjunto de classes é feito a partir daquela que possui o maior número de instâncias. Dessa forma, considerando $i$ a quantidade de instâncias desejadas e $j$ a quantidade de instâncias da classe desbalanceada, o cálculo para a porcentagem de aumento pode ser visualizado em (2.6):

$$Porcentagem = \left(\frac{i}{j} x\ 100\right) -\ 100 \tag{2.6}$$

## 3.7 TREINAMENTO E TESTE

As 6 bases de dados balanceadas foram submetidas ao treinamento de classificação com validação cruzada de 10 *folds* e 30 repetições pelo software *Weka*. A relação dos algoritmos e configurações específicas aplicadas pode ser visualizada na Tabela 2. Logo após, a partir da planilha de desempenho gerada, foram extraídas a média e o desvio padrão dos indicadores estatísticos mencionados na fundamentação teórica: acurácia, índice kappa, sensibilidade, especificidade e AUC-ROC. O melhor classificador, baseado nas métricas citadas, foi implementado nas bases de teste para analisar o desempenho do algoritmo.

Tabela 2 – Algoritmos de classificação utilizados no trabalho

| ALGORITMO | CONFIGURAÇÃO |
|---|---|
| Random Forest | 10 interações |
| | 20 interações |
| | 50 interações |
| | 100 interações |
| | 150 interações |
| | 200 interações |
| | 250 interações |
| | 300 interações |
| SVM | Função linear |
| | Funçao polinomial 2 |
| | Função polinomial 3 |
| | Função polinomial 4 |
| | RBF (gamma = 0.01) |
| | RBF (gamma = 0.25) |
| | RBF (gamma = 0.5) |
| J48 | batchSize = 100 |
| Bayes Network | batchSize = 100 |
| Naive Bayes | batchSize = 100 |

Fonte: O autor (2024)

## 3.8 ANÁLISE DA RELEVÂNCIA DE BIOMARCADORES NO CÂNCER DE MAMA

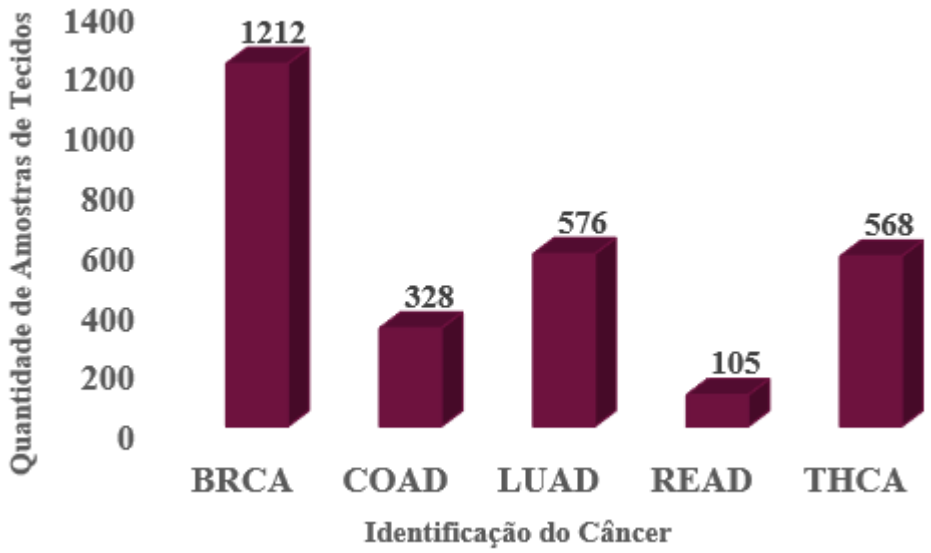
Após as etapas de treinamento e teste das subamostras baseadas no *Random Forest* (n=10), foram feitas análises, na base de dados primordial, com números de interações maiores (n=20, n=50 e n=100), visando verificar possíveis mudanças na relevância dos genes BRCA1 e BRCA2, comuns no diagnóstico do câncer de mama especificamente.

# 4 RESULTADOS E DISCUSSÕES

## 4.1 A BASE DE DADOS DE TRABALHO

A partir da comparação com o gráfico de incidência de câncer em mulheres no país (Figura 1) e dos dados disponibilizados pela *Broad Institute*, os 5 tipos de câncer escolhidos para a amostra de dados foram: Carcinoma invasivo de mama (BRCA), Adenocarcinoma de cólon (COAD), Adenocarcinoma pulmonar (LUAD), Adenocarcinoma de reto (READ) e Carcinoma de tireoide (THCA). A distribuição específica dos tecidos pelos tipos de câncer selecionados pode ser visualizada na Figura 6.

Figura 6 – Distribuição das amostras de tecidos por tipo de câncer




Fonte: Broad GDAC Firehose (2016)

Nota: Tabela elaborada com base nos dados informados pelo *Broad GDAC Firehose*.

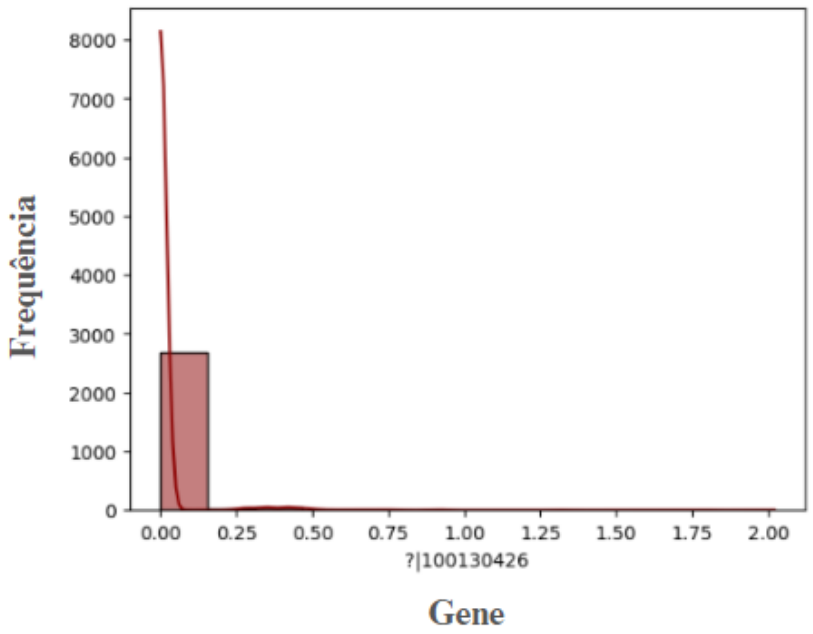
A base de dados de trabalho compreende 2789 amostras, sendo aproximadamente 26,7% da base de dados total disponibilizada pelo Instituto.

## 4.2 ANÁLISE ESTATÍSTICA DESCRITIVA

A distribuição estatística descritiva da base de dados de trabalho foi baseada na expressão dos genes de referência nas amostras de câncer. A partir disso, foram visualizados 4 agrupamentos principais:

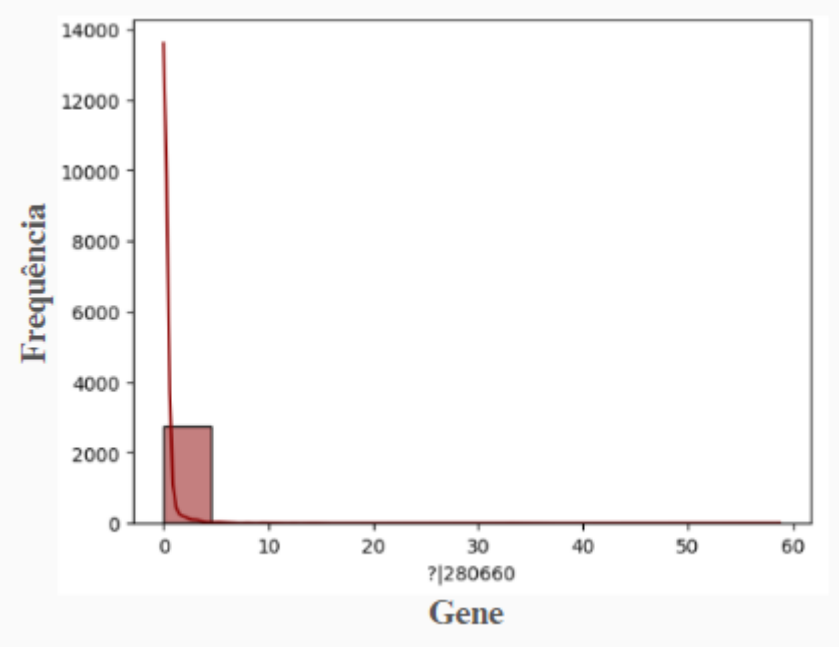
1. Agrupamento 1 - Genes com pouca expressão 1: Genes com expressão nula na maioria dos tecidos, mas quando são expressadas possuem valores próximos de zero e da média, como pode ser visualizado no Gráfico 1.
2. Agrupamento 2 - Genes com pouca expressão 2: Genes com expressão nula na maioria dos tecidos, mas quando são expressadas possuem valores mais distantes do zero e da média, como pode ser verificado no Gráfico 2.
3. Agrupamento 3 - Genes com muitas expressões 1: Genes com expressão significativa na maioria dos tecidos, porém há expressão nula em algumas amostras dos tipos de câncer, como pode ser visualizado no Gráfico 3.
4. Agrupamento 4 - Genes com muitas expressões 2: Genes com expressão significativa na maioria dos tecidos e sem expressão nula, como pode ser visualizado no Gráfico 4.

Gráfico 1: Distribuição de um exemplo de gene do agrupamento 1



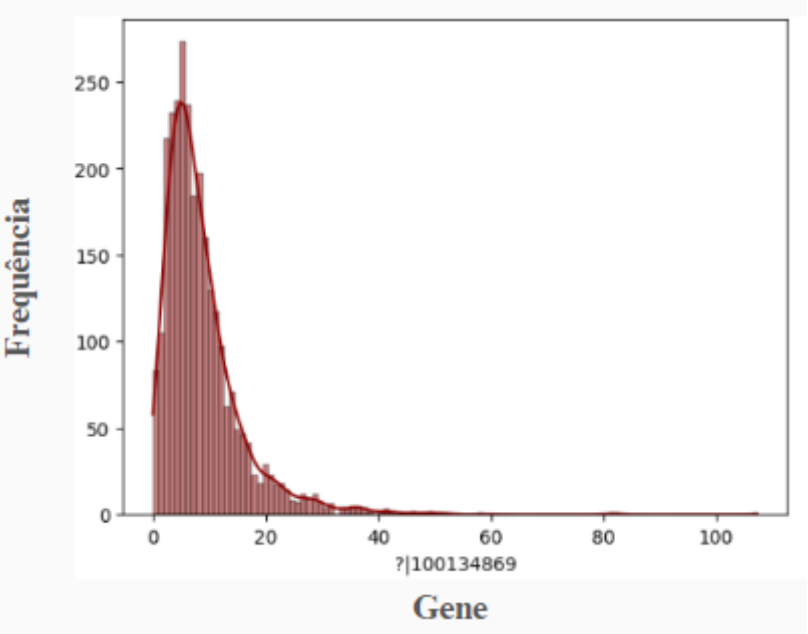

Fonte: O autor (2024)

Gráfico 2: Distribuição de um exemplo de gene do agrupamento 2



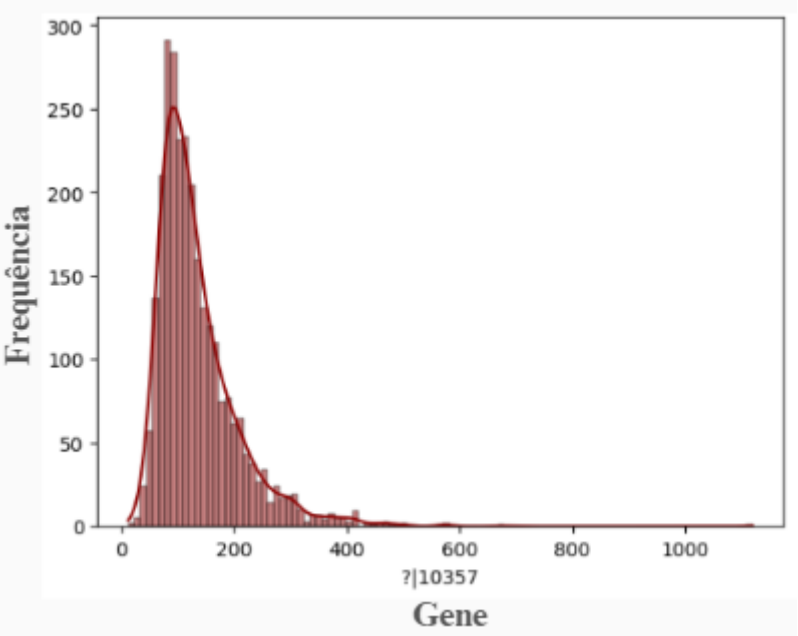

Fonte: O autor (2024)

Gráfico 3: Distribuição de um exemplo de gene do agrupamento 3




Fonte: O autor (2024)

Gráfico 4: Distribuição de um exemplo de gene do agrupamento 4




Fonte: O autor (2024)

## 4.3 SUBAMOSTRAGENS DA BASE DE DADOS DE TRABALHO

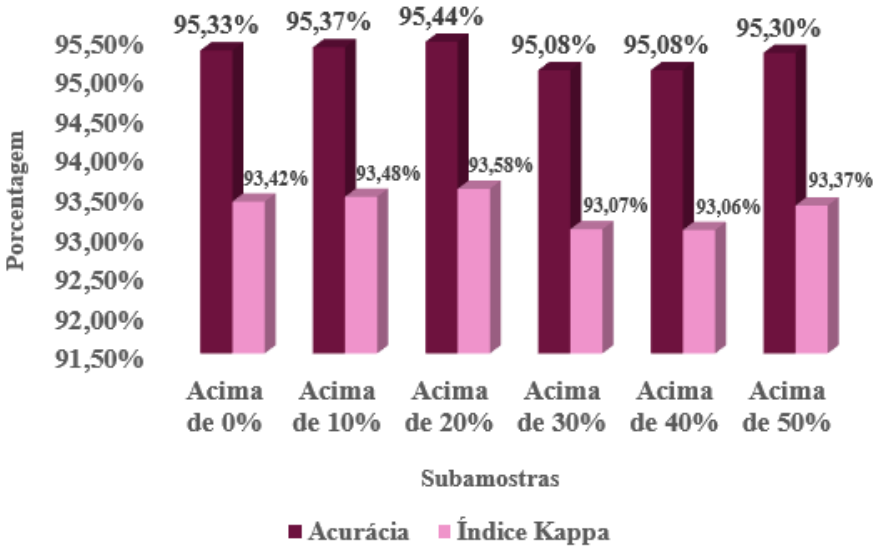
A partir do classificador *Random Forest* (n=10), foi possível considerar a relevância de todos os genes para extrair 6 subamostras: relevâncias acima de 0%, 10%, 20%, 30%, 40% e 50%. A distribuição de genes para cada subamostra pode ser visualizada na Tabela 3. Também foi possível extrair a acurácia e o índice kappa iniciais das subamostras, como pode ser verificado no Gráfico 5.

Tabela 3 – Distribuição dos atributos (genes) nas subamostras da base de dados

| SUBAMOSTRA | QUANTIDADE DE ATRIBUTOS | QUANTIDADE DE INSTÂNCIAS | PORCENTAGEM DE DIMINUIÇÃO DE ATRIBUTOS |
|---|---|---|---|
| Acima de 0% | 1330 | 2789 | 87,27% |
| Acima de 10% | 1194 | 2789 | 10,22% |
| Acima de 20% | 1028 | 2789 | 13,90% |
| Acima de 30% | 849 | 2789 | 17,41% |
| Acima de 40% | 719 | 2789 | 15,31% |
| Acima de 50% | 588 | 2789 | 18,22% |

Fonte: O autor (2024)

Gráfico 5 – Análise de aprendizado do modelo *Random Forest* (n = 10)




Fonte: O autor (2024)

## 4.4 PRÉ-TREINAMENTO DAS SUBAMOSTRAS

Inicialmente, as subamostras foram separadas, pelo método de *Resample*, em bases de treinamento (70%) e teste (30%). A nova distribuição dos dados das subamostras e das 5 classes podem ser visualizadas nas Tabelas 4 e 5.

Tabela 4 – Nova distribuição das instâncias por base de dados de treinamento e teste

| SUBAMOSTRA | QUANTIDADE DE ATRIBUTOS | QUANTIDADE DE INSTÂNCIAS - TREINAMENTO | QUANTIDADE DE INSTÂNCIAS - TESTE |
|---|---|---|---|
| Acima de 0% | 1330 | 1950 | 587 |
| Acima de 10% | 1194 | 1950 | 587 |
| Acima de 20% | 1028 | 1950 | 587 |
| Acima de 30% | 849 | 1950 | 587 |
| Acima de 40% | 719 | 1950 | 587 |
| Acima de 50% | 588 | 1950 | 587 |

Fonte: O autor (2024)

Tabela 5 – Distribuição das instâncias de treinamento e teste por classe

| CLASSE | QUANTIDADE DE INSTÂNCIAS - TREINAMENTO | QUANTIDADE DE INSTÂNCIAS - TESTE |
|---|---|---|
| BRCA | 848 | 255 |
| COAD | 229 | 69 |
| LUAD | 403 | 121 |
| READ | 73 | 22 |
| THCA | 397 | 120 |

Fonte: O autor (2024)

Com relação ao balanceamento das bases de treinamento, utilizando o método *SMOTE*, a quantidade de instâncias foi aumentada seguindo a classe BRCA. A porcentagem de aumento para as 4 classes pode ser visualizada na Tabela 6.

Tabela 6 – Porcentagem de aumento das classes balanceadas

| CLASSE | PORCENTAGEM DE AUMENTO |
|---|---|
| COAD | 270% |
| LUAD | 110% |
| READ | 1062% |
| THCA | 114% |

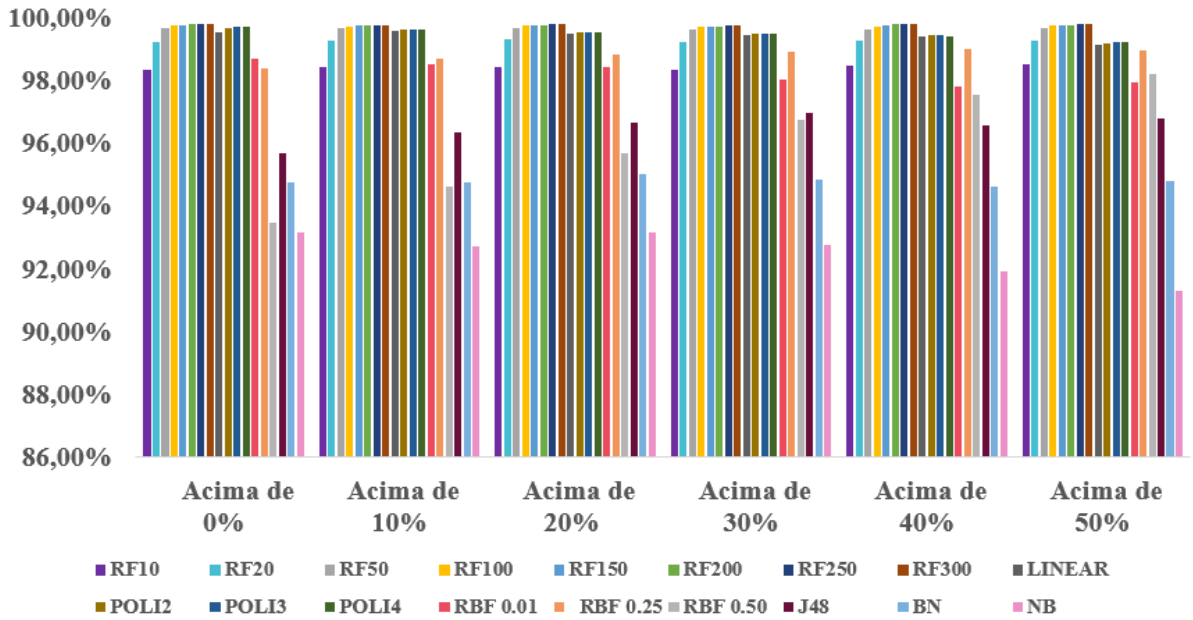
Fonte: O autor (2024)

## 4.5 TREINAMENTO E TESTE

### 4.5.1 Treinamento

A partir de 10 *folds* e 30 repetições, cada algoritmo descrito na Tabela 2 foi treinado na base de dados de treinamento. O desempenho dos classificadores pode ser visualizado no Gráfico 6 e nas Tabelas 7, 8, 9, 10, 11 e 12. Todas as métricas foram incorporadas com 5 números significativos após a vírgula para melhor análise de escolha do algoritmo.

Gráfico 6 – Desempenho dos algoritmos de treinamento nas subamostras




Fonte: O autor (2024)

Tabela 7 – Desempenho do treinamento para genes acima de 0% de relevância

**ACIMA DE 0%**

| CLASSIFICADORES | | ACURÁCIA | | ÍNDICE KAPPA | | SENSIBILIDADE | | ESPECIFICIDADE | | AUC-ROC | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| ALGORITMO | CONFIGURAÇÃO | MÉDIA | DESVIO PADRÃO | MÉDIA | DESVIO PADRÃO | MÉDIA | DESVIO PADRÃO | MÉDIA | DESVIO PADRÃO | MÉDIA | DESVIO PADRÃO |
| Random Forest | 10 interações | 98,33571 | 0,70254 | 0,97920 | 0,00878 | 0,98318 | 0,01464 | 0,98670 | 0,00692 | 0,99875 | 0,00102 |
| | 20 interações | 99,24809 | 0,44838 | 0,99060 | 0,00560 | 0,99127 | 0,01073 | 0,99499 | 0,00420 | 0,99973 | 0,00032 |
| | 50 interações | 99,65235 | 0,28262 | 0,99565 | 0,00353 | 0,99568 | 0,00746 | 0,99820 | 0,00240 | 0,99995 | 0,00009 |
| | 100 interações | 99,74359 | 0,23469 | 0,99679 | 0,00293 | 0,99701 | 0,00573 | 0,99882 | 0,00187 | 0,99998 | 0,00005 |
| | 150 interações | 99,78135 | 0,22019 | 0,99727 | 0,00275 | 0,99780 | 0,00490 | 0,99895 | 0,00186 | 0,99999 | 0,00003 |
| | 200 interações | 99,79707 | 0,20245 | 0,99746 | 0,00253 | 0,99823 | 0,00443 | 0,99901 | 0,00166 | 0,99999 | 0,00003 |
| | 250 interações | 99,79707 | 0,19874 | 0,99746 | 0,00248 | 0,99811 | 0,00443 | 0,99900 | 0,00168 | 0,99999 | 0,00003 |
| | **300 interações** | **99,79943** | **0,19504** | **0,99749** | **0,00244** | **0,99815** | **0,00440** | **0,99901** | **0,00166** | **0,99999** | **0,00003** |
| SVM | Linear | 99,54461 | 0,30316 | 0,99431 | 0,00379 | 0,98495 | 0,01207 | 0,99875 | 0,00188 | 0,99751 | 0,00181 |
| | Polinomial 2 | 99,68146 | 0,26820 | 0,99602 | 0,00335 | 0,98911 | 0,01084 | 0,99934 | 0,00141 | 0,99833 | 0,00156 |
| | Polinomial 3 | 99,72471 | 0,23464 | 0,99656 | 0,00293 | 0,99065 | 0,00986 | 0,99949 | 0,00117 | 0,99861 | 0,00136 |
| | Polinomial 4 | 99,72551 | 0,23671 | 0,99657 | 0,00296 | 0,99140 | 0,00927 | 0,99946 | 0,00124 | 0,99867 | 0,00133 |
| | RBF 0.01 | 98,71094 | 0,53650 | 0,98389 | 0,00671 | 0,97822 | 0,01629 | 0,99044 | 0,00506 | 0,99253 | 0,00321 |
| | RBF 0.25 | 98,37108 | 0,62424 | 0,97964 | 0,00780 | 0,98384 | 0,01354 | 0,99979 | 0,00075 | 0,99674 | 0,00338 |
| | RBF 0.50 | 93,49294 | 1,20223 | 0,91866 | 0,01503 | 0,96144 | 0,02064 | 0,99997 | 0,00029 | 0,98809 | 0,00740 |
| J48 | batchSize = 100 | 95,69292 | 1,44824 | 0,94616 | 0,01810 | 0,94618 | 0,03032 | 0,96719 | 0,01441 | 0,96772 | 0,01920 |
| Bayes Network | batchSize = 100 | 94,76091 | 1,06160 | 0,93451 | 0,01327 | 0,96517 | 0,00979 | 0,96517 | 0,00979 | 0,98652 | 0,00471 |
| Naive Bayes | batchSize = 100 | 93,18322 | 1,23424 | 0,91479 | 0,01543 | 0,90445 | 0,03364 | 0,94560 | 0,01184 | 0,97128 | 0,00784 |

Fonte: O autor (2024)

Tabela 8 – Desempenho do treinamento para genes acima de 10% de relevância

**ACIMA DE 10%**

| CLASSIFICADORES | | ACURÁCIA | | ÍNDICE KAPPA | | SENSIBILIDADE | | ESPECIFICIDADE | | AUC-ROC | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| ALGORITMO | CONFIGURAÇÃO | MÉDIA | DESVIO PADRÃO | MÉDIA | DESVIO PADRÃO | MÉDIA | DESVIO PADRÃO | MÉDIA | DESVIO PADRÃO | MÉDIA | DESVIO PADRÃO |
| Random Forest | 10 interações | 98,41671 | 0,61898 | 0,98021 | 0,00774 | 0,98463 | 0,01393 | 0,98730 | 0,00641 | 0,99890 | 0,00087 |
| | 20 interações | 99,25280 | 0,46025 | 0,99066 | 0,00575 | 0,98970 | 0,01076 | 0,99537 | 0,00410 | 0,99974 | 0,00031 |
| | 50 interações | 99,65628 | 0,27914 | 0,99570 | 0,00349 | 0,99544 | 0,00706 | 0,99831 | 0,00232 | 0,99995 | 0,00009 |
| | 100 interações | 99,72156 | 0,25248 | 0,99652 | 0,00316 | 0,99607 | 0,00663 | 0,99887 | 0,00189 | 0,99998 | 0,00005 |
| | 150 interações | 99,73965 | 0,24858 | 0,99675 | 0,00311 | 0,99631 | 0,00662 | 0,99905 | 0,00174 | 0,99998 | 0,00004 |
| | 200 interações | 99,74437 | 0,24294 | 0,99680 | 0,00304 | 0,99603 | 0,00671 | 0,99919 | 0,00158 | 0,99999 | 0,00004 |
| | 250 interações | 99,74358 | 0,23937 | 0,99679 | 0,00299 | 0,99572 | 0,00686 | 0,99874 | 0,00193 | 0,99998 | 0,00005 |
| | **300 interações** | **99,76246** | **0,23318** | **0,99703** | **0,00291** | **0,99631** | **0,00634** | **0,99925** | **0,00153** | **0,99999** | **0,00004** |
| SVM | Linear | 99,59022 | 0,30091 | 0,99488 | 0,00376 | 0,98702 | 0,01212 | 0,99893 | 0,00173 | 0,99788 | 0,00172 |
| | Polinomial 2 | 99,63898 | 0,27250 | 0,99549 | 0,00341 | 0,98808 | 0,01091 | 0,99920 | 0,00144 | 0,99814 | 0,00153 |
| | Polinomial 3 | 99,64134 | 0,28196 | 0,99552 | 0,00352 | 0,98812 | 0,01106 | 0,99928 | 0,00146 | 0,99825 | 0,00155 |
| | Polinomial 4 | 99,63427 | 0,27164 | 0,99543 | 0,00340 | 0,98828 | 0,01062 | 0,99944 | 0,00123 | 0,99835 | 0,00147 |
| | RBF 0.01 | 98,54106 | 0,60687 | 0,98176 | 0,00759 | 0,97181 | 0,01858 | 0,99066 | 0,00537 | 0,99189 | 0,00353 |
| | RBF 0.25 | 98,69751 | 0,55812 | 0,98372 | 0,00698 | 0,98518 | 0,01248 | 0,99971 | 0,00087 | 0,99733 | 0,00274 |
| | RBF 0.50 | 94,62458 | 1,04881 | 0,93281 | 0,01311 | 0,95638 | 0,01956 | 0,99997 | 0,00029 | 0,98738 | 0,00029 |
| J48 | batchSize = 100 | 96,34559 | 1,35675 | 0,95432 | 0,01696 | 0,94635 | 0,03494 | 0,97325 | 0,01228 | 0,94635 | 0,02071 |
| Bayes Network | batchSize = 100 | 94,76091 | 1,35675 | 0,93451 | 0,01327 | 0,90633 | 0,03128 | 0,96517 | 0,00979 | 0,98652 | 0,00471 |
| Naive Bayes | batchSize = 100 | 92,70886 | 1,23944 | 0,90886 | 0,01549 | 0,88479 | 0,03517 | 0,94421 | 0,01140 | 0,96963 | 0,00824 |

Fonte: O autor (2024)

Tabela 9 – Desempenho do treinamento para genes acima de 20% de relevância

| ACIMA DE 20% | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| CLASSIFICADORES | | ACURÁCIA | | ÍNDICE KAPPA | | SENSIBILIDADE | | ESPECIFICIDADE | | AUC-ROC | |
| ALGORITMO | CONFIGURAÇÃO | MÉDIA | DESVIO PADRÃO | MÉDIA | DESVIO PADRÃO | MÉDIA | DESVIO PADRÃO | MÉDIA | DESVIO PADRÃO | MÉDIA | DESVIO PADRÃO |
| Random Forest | 10 interações | 98,42685 | 0,65214 | 0,98034 | 0,00815 | 0,98376 | 0,01352 | 0,98785 | 0,00650 | 0,99893 | 0,00090 |
| | 20 interações | 99,31334 | 0,42690 | 0,99142 | 0,00534 | 0,99057 | 0,01087 | 0,99548 | 0,00409 | 0,99976 | 0,00031 |
| | 50 interações | 99,66179 | 0,26380 | 0,99577 | 0,00330 | 0,99454 | 0,00767 | 0,99840 | 0,00199 | 0,99994 | 0,00011 |
| | 100 interações | 99,75225 | 0,23724 | 0,99690 | 0,00297 | 0,99552 | 0,00736 | 0,99896 | 0,00169 | 0,99998 | 0,00005 |
| | 150 interações | 99,76483 | 0,23036 | 0,99706 | 0,00288 | 0,99552 | 0,00729 | 0,99909 | 0,00167 | 0,99998 | 0,00004 |
| | 200 interações | 99,77427 | 0,23175 | 0,99718 | 0,00290 | 0,99596 | 0,00700 | 0,99911 | 0,00170 | 0,99998 | 0,00004 |
| | **250 interações** | **99,78528** | **0,21254** | **0,99732** | **0,00266** | **0,99584** | **0,00662** | **0,99919** | **0,00161** | **0,99999** | **0,00003** |
| | 300 interações | 99,78371 | 0,21871 | 0,99730 | 0,00273 | 0,99580 | 0,00691 | 0,99920 | 0,00165 | 0,99999 | 0,00003 |
| SVM | Linear | 99,51157 | 0,34348 | 0,99389 | 0,00429 | 0,98416 | 0,01374 | 0,99833 | 0,00216 | 0,99724 | 0,00207 |
| | Polinomial 2 | 99,53123 | 0,33311 | 0,99414 | 0,00416 | 0,98478 | 0,01319 | 0,99829 | 0,00227 | 0,99740 | 0,00196 |
| | Polinomial 3 | 99,54617 | 0,32900 | 0,99433 | 0,00411 | 0,98463 | 0,01354 | 0,99853 | 0,00201 | 0,99743 | 0,00196 |
| | Polinomial 4 | 99,53673 | 0,32824 | 0,99421 | 0,00410 | 0,98349 | 0,01385 | 0,99892 | 0,00179 | 0,99746 | 0,00196 |
| | RBF 0.01 | 98,41124 | 0,62152 | 0,98014 | 0,00777 | 0,96961 | 0,01835 | 0,98945 | 0,00546 | 0,99101 | 0,00367 |
| | RBF 0.25 | 98,82964 | 0,53606 | 0,98537 | 0,00670 | 0,98094 | 0,01529 | 0,99965 | 0,00099 | 0,99684 | 0,00296 |
| | RBF 0.50 | 95,67169 | 1,05589 | 0,94590 | 0,01320 | 0,96049 | 0,02085 | 0,99971 | 0,00089 | 0,98937 | 0,00738 |
| J48 | batchSize = 100 | 96,67931 | 1,14718 | 0,95849 | 0,01434 | 0,96128 | 0,02609 | 0,97401 | 0,01120 | 0,98002 | 0,01576 |
| Bayes Network | batchSize = 100 | 95,03465 | 0,98193 | 0,93793 | 0,01227 | 0,91065 | 0,03160 | 0,96864 | 0,00866 | 0,98790 | 0,00424 |
| Naive Bayes | batchSize = 100 | 93,18322 | 1,23424 | 0,91479 | 0,01543 | 0,90445 | 0,03364 | 0,94560 | 0,01184 | 0,97128 | 0,00784 |

Fonte: O autor (2024)

Tabela 10 – Desempenho do treinamento para genes acima de 30% de relevância

| ACIMA DE 30% | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| CLASSIFICADORES | | ACURÁCIA | | ÍNDICE KAPPA | | SENSIBILIDADE | | ESPECIFICIDADE | | AUC-ROC | |
| ALGORITMO | CONFIGURAÇÃO | MÉDIA | DESVIO PADRÃO | MÉDIA | DESVIO PADRÃO | MÉDIA | DESVIO PADRÃO | MÉDIA | DESVIO PADRÃO | MÉDIA | DESVIO PADRÃO |
| Random Forest | 10 interações | 98,35305 | 0,63254 | 0,97941 | 0,00791 | 0,98345 | 0,01500 | 0,98672 | 0,00644 | 0,99874 | 0,00105 |
| | 20 interações | 99,23785 | 0,41460 | 0,99047 | 0,00518 | 0,99096 | 0,01055 | 0,99449 | 0,00397 | 0,99971 | 0,00033 |
| | 50 interações | 99,61067 | 0,30640 | 0,99513 | 0,00383 | 0,99473 | 0,00760 | 0,99770 | 0,00265 | 0,99993 | 0,00012 |
| | 100 interações | 99,71213 | 0,26155 | 0,99640 | 0,00327 | 0,99576 | 0,00692 | 0,99843 | 0,00208 | 0,99996 | 0,00007 |
| | 150 interações | 99,73730 | 0,24105 | 0,99672 | 0,00301 | 0,99868 | 0,00187 | 0,99519 | 0,00500 | 0,99997 | 0,00006 |
| | 200 interações | 99,73258 | 0,24960 | 0,99666 | 0,00312 | 0,99540 | 0,00693 | 0,99870 | 0,00193 | 0,99997 | 0,00006 |
| | **250 interações** | **99,74358** | **0,23937** | **0,99679** | **0,00299** | **0,99572** | **0,00686** | **0,99874** | **0,00193** | **0,99998** | **0,00005** |
| | 300 interações | 99,73809 | 0,24461 | 0,99673 | 0,00306 | 0,99528 | 0,00708 | 0,99877 | 0,00183 | 0,99998 | 0,00006 |
| SVM | Linear | 99,43448 | 0,33141 | 0,99293 | 0,00414 | 0,97963 | 0,01425 | 0,99862 | 0,00216 | 0,99682 | 0,00204 |
| | Polinomial 2 | 99,50920 | 0,30455 | 0,99386 | 0,00381 | 0,98211 | 0,01350 | 0,99893 | 0,00183 | 0,99729 | 0,00187 |
| | Polinomial 3 | 99,51313 | 0,30201 | 0,99391 | 0,00378 | 0,98235 | 0,01337 | 0,99893 | 0,00180 | 0,99733 | 0,00185 |
| | Polinomial 4 | 99,47854 | 0,33417 | 0,99348 | 0,00418 | 0,98121 | 0,01426 | 0,99892 | 0,00191 | 0,99716 | 0,00204 |
| | RBF 0.01 | 98,05341 | 0,67691 | 0,97567 | 0,00846 | 0,97127 | 0,01708 | 0,98427 | 0,00705 | 0,98863 | 0,00411 |
| | RBF 0.25 | 98,90279 | 0,51604 | 0,98628 | 0,00645 | 0,97822 | 0,01495 | 0,99970 | 0,00090 | 0,99651 | 0,00294 |
| | RBF 0.50 | 96,76655 | 0,88752 | 0,95958 | 0,01109 | 0,96490 | 0,01903 | 0,99965 | 0,00096 | 0,99342 | 0,00474 |
| J48 | batchSize = 100 | 96,95291 | 0,83455 | 0,96191 | 0,01043 | 0,97504 | 0,02034 | 0,97269 | 0,00865 | 0,98984 | 0,00837 |
| Bayes Network | batchSize = 100 | 94,85449 | 1,02977 | 0,93568 | 0,01287 | 0,90962 | 0,03145 | 0,96567 | 0,00926 | 0,98718 | 0,00456 |
| Naive Bayes | batchSize = 100 | 92,75682 | 1,27806 | 0,90946 | 0,01598 | 0,87303 | 0,03556 | 0,94793 | 0,01158 | 0,97089 | 0,00784 |

Fonte: O autor (2024)

Tabela 11 – Desempenho do treinamento para genes acima de 40% de relevância

**ACIMA DE 40%**

| CLASSIFICADORES | | ACURÁCIA | | ÍNDICE KAPPA | | SENSIBILIDADE | | ESPECIFICIDADE | | AUC-ROC | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| ALGORITMO | CONFIGURAÇÃO | MÉDIA | DESVIO PADRÃO | MÉDIA | DESVIO PADRÃO | MÉDIA | DESVIO PADRÃO | MÉDIA | DESVIO PADRÃO | MÉDIA | DESVIO PADRÃO |
| Random Forest | 10 interações | 98,48437 | 0,65315 | 0,98105 | 0,00816 | 0,98397 | 0,01479 | 0,98828 | 0,00643 | 0,99893 | 0,00098 |
| | 20 interações | 99,29133 | 0,44490 | 0,99114 | 0,00556 | 0,99139 | 0,01037 | 0,99516 | 0,00389 | 0,99974 | 0,00034 |
| | 50 interações | 99,64763 | 0,26426 | 99,64763 | 0,00330 | 0,99477 | 0,00783 | 0,99819 | 0,00219 | 0,99994 | 0,00012 |
| | 100 interações | 99,73729 | 0,24644 | 0,99672 | 0,00308 | 0,99568 | 0,00700 | 0,99889 | 0,00178 | 0,99997 | 0,00006 |
| | 150 interações | 99,77898 | 0,22497 | 0,99724 | 0,00281 | 0,99603 | 0,00657 | 0,99925 | 0,00150 | 0,99998 | 0,00004 |
| | 200 interações | 99,78920 | 0,21687 | 0,99737 | 0,00271 | 0,99572 | 0,00679 | 0,99928 | 0,00142 | 0,99998 | 0,00004 |
| | 250 interações | 99,79865 | 0,21642 | 0,99748 | 0,00271 | 0,99560 | 0,00708 | 0,99934 | 0,00141 | 0,99998 | 0,00004 |
| | **300 interações** | **99,80022** | **0,20649** | **0,99750** | **0,00258** | **0,99560** | **0,00661** | **0,99940** | **0,00135** | **0,99998** | **0,00004** |
| SVM | Linear | 99,40223 | 0,34919 | 0,99253 | 0,00436 | 0,97767 | 0,97771 | 0,99857 | 0,00207 | 0,99656 | 0,00208 |
| | Polinomial 2 | 99,45965 | 0,32461 | 0,99325 | 0,00406 | 0,97905 | 0,01488 | 0,99900 | 0,00159 | 0,99694 | 0,00193 |
| | Polinomial 3 | 99,45650 | 0,33692 | 0,99321 | 0,00421 | 0,97991 | 0,01502 | 0,99899 | 0,00165 | 0,99704 | 0,00196 |
| | Polinomial 4 | 99,42111 | 0,33160 | 0,99276 | 0,00414 | 0,97932 | 0,01507 | 0,99898 | 0,00165 | 0,99697 | 0,00193 |
| | RBF 0.01 | 97,82926 | 0,72289 | 0,97287 | 0,00904 | 0,95514 | 0,02138 | 0,98593 | 0,00668 | 0,98750 | 0,00426 |
| | RBF 0.25 | 98,98852 | 0,46183 | 0,98736 | 0,00577 | 0,97551 | 0,01575 | 0,99939 | 0,00133 | 0,99615 | 0,00278 |
| | RBF 0.50 | 97,53582 | 0,75054 | 0,96920 | 0,00938 | 0,96124 | 0,02028 | 0,99924 | 0,00142 | 0,99275 | 0,00441 |
| J48 | batchSize = 100 | 96,58483 | 1,07457 | 0,95731 | 0,01343 | 0,97095 | 0,02848 | 0,96936 | 0,01101 | 0,98523 | 0,01626 |
| Bayes Network | batchSize = 100 | 94,62721 | 1,10517 | 0,93284 | 0,01381 | 0,90015 | 0,03637 | 0,96583 | 0,00928 | 0,98880 | 0,00426 |
| Naive Bayes | batchSize = 100 | 91,91051 | 1,23235 | 0,89888 | 0,01540 | 0,85227 | 0,03768 | 0,94366 | 0,01143 | 0,96976 | 0,00748 |

Fonte: O autor (2024)

Tabela 12 – Desempenho do treinamento para genes acima de 50% de relevância

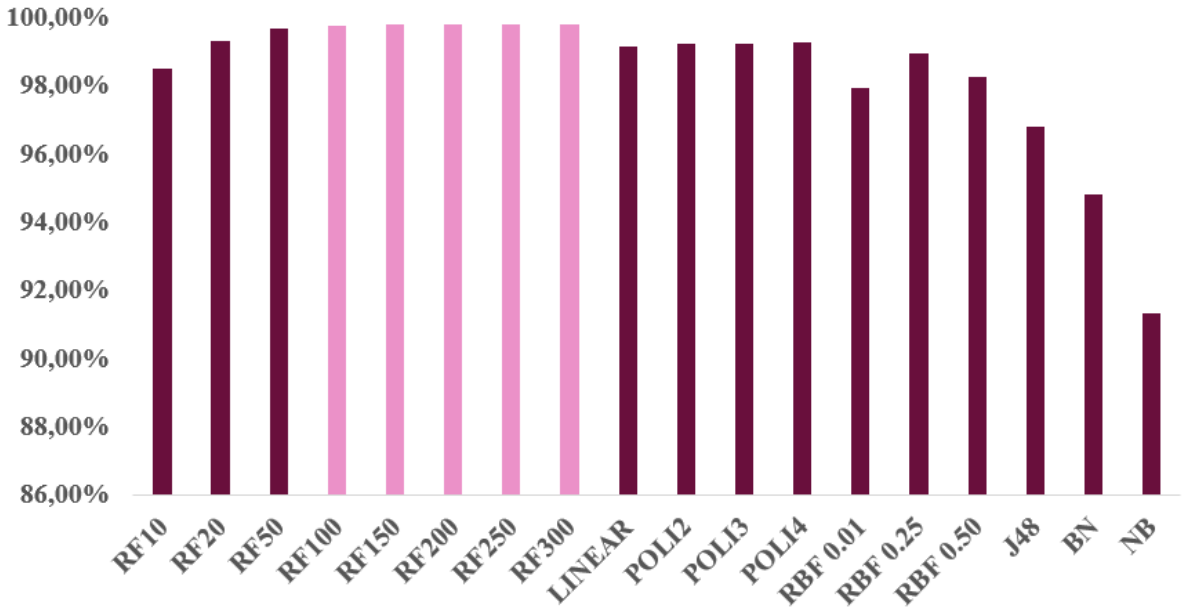
**ACIMA DE 50%**

| CLASSIFICADORES | | ACURÁCIA | | ÍNDICE KAPPA | | SENSIBILIDADE | | ESPECIFICIDADE | | AUC-ROC | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| ALGORITMO | CONFIGURAÇÃO | MÉDIA | DESVIO PADRÃO | MÉDIA | DESVIO PADRÃO | MÉDIA | DESVIO PADRÃO | MÉDIA | DESVIO PADRÃO | MÉDIA | DESVIO PADRÃO |
| Random Forest | 10 interações | 98,50717 | 0,63511 | 0,98134 | 0,00793 | 0,98554 | 0,01265 | 0,98833 | 0,00634 | 0,99896 | 0,00096 |
| | 20 interações | 99,29056 | 0,40979 | 0,99113 | 0,00512 | 0,99115 | 0,01008 | 0,99532 | 0,00376 | 0,99976 | 0,00029 |
| | 50 interações | 99,66023 | 0,27624 | 0,99575 | 0,00345 | 0,99485 | 0,00872 | 0,99824 | 0,00212 | 0,99995 | 0,00010 |
| | 100 interações | 99,75067 | 0,24027 | 0,99688 | 0,00300 | 0,99611 | 0,00696 | 0,99889 | 0,00173 | 0,99997 | 0,00005 |
| | 150 interações | 99,77427 | 0,21935 | 0,99718 | 0,00274 | 0,99631 | 0,00627 | 0,99911 | 0,00154 | 0,99998 | 0,00004 |
| | 200 interações | 99,77741 | 0,21228 | 0,99722 | 0,00265 | 0,99639 | 0,00624 | 0,99917 | 0,00151 | 0,99999 | 0,00003 |
| | 250 interações | 99,78528 | 0,21855 | 0,99732 | 0,00273 | 0,99670 | 0,00611 | 0,99924 | 0,00148 | 0,99999 | 0,00003 |
| | **300 interações** | **99,79157** | **0,22124** | **0,99739** | **0,00277** | **0,99690** | **0,00571** | **0,99927** | **0,00154** | **0,99999** | **0,00003** |
| SVM | Linear | 99,13877 | 0,39457 | 0,98923 | 0,00493 | 0,97056 | 0,01683 | 0,99776 | 0,00246 | 0,99525 | 0,00234 |
| | Polinomial 2 | 99,20089 | 0,37941 | 0,99001 | 0,00474 | 0,97331 | 0,01649 | 0,99787 | 0,00252 | 0,99566 | 0,00231 |
| | Polinomial 3 | 99,23707 | 0,38585 | 0,99046 | 0,00482 | 0,97319 | 0,01665 | 0,99833 | 0,00235 | 0,99588 | 0,00231 |
| | Polinomial 4 | 99,24337 | 0,37400 | 0,99054 | 0,00467 | 0,97287 | 0,01656 | 0,99853 | 0,00212 | 0,99594 | 0,00224 |
| | RBF 0.01 | 97,93539 | 0,67389 | 0,97419 | 0,00842 | 0,95051 | 0,02359 | 0,98910 | 0,00550 | 0,98850 | 0,00387 |
| | RBF 0.25 | 98,94447 | 0,47839 | 0,98681 | 0,00598 | 0,96820 | 0,01788 | 0,99969 | 0,00091 | 0,99566 | 0,00246 |
| | RBF 0.50 | 98,23190 | 0,67217 | 0,97790 | 0,00840 | 0,96568 | 0,01902 | 0,99958 | 0,00112 | 0,99444 | 0,00404 |
| J48 | batchSize = 100 | 96,79092 | 0,81927 | 0,95989 | 0,01024 | 0,98507 | 0,01797 | 0,98507 | 0,01797 | 0,99253 | 0,00793 |
| Bayes Network | batchSize = 100 | 94,81758 | 0,97769 | 0,93522 | 0,01222 | 0,91128 | 0,03335 | 0,96564 | 0,00872 | 0,98950 | 0,00387 |
| Naive Bayes | batchSize = 100 | 91,29780 | 1,26666 | 0,89122 | 0,01583 | 0,85508 | 0,03649 | 0,93631 | 0,01276 | 0,96974 | 0,00751 |

Fonte: O autor (2024)

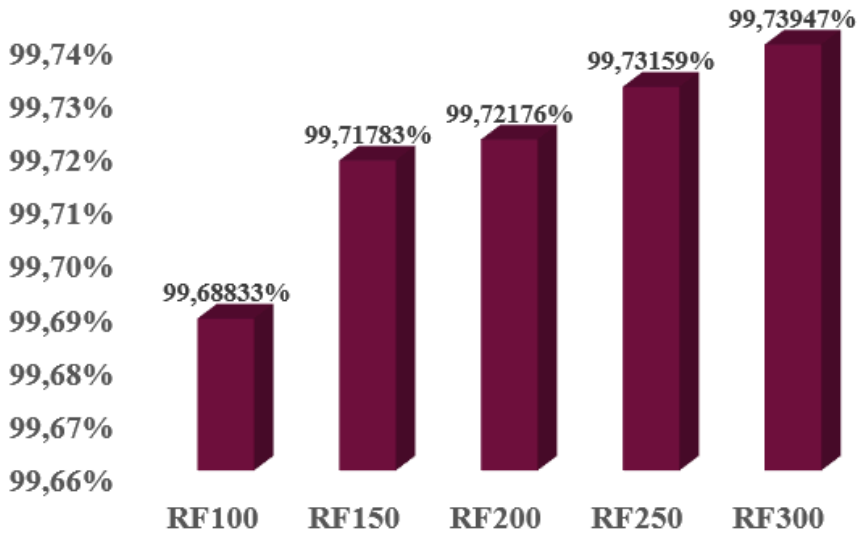
A partir do Gráfico 6, é possível visualizar que o algoritmo *Random Forest* entre 100 e 300 interações obteve um desempenho bastante próximo com diferenças na escala de centésimos e milésimos na acurácia, além de valores iguais em outros indicadores, como na AUC-ROC. Ademais, é possível verificar, graficamente, a diferença ínfima entre os indicadores estatísticos na base de dados com genes acima de 50% nos Gráficos 7, 8 e 9.

Gráfico 7 – Acurácia de treinamento para genes acima de 50% de relevância



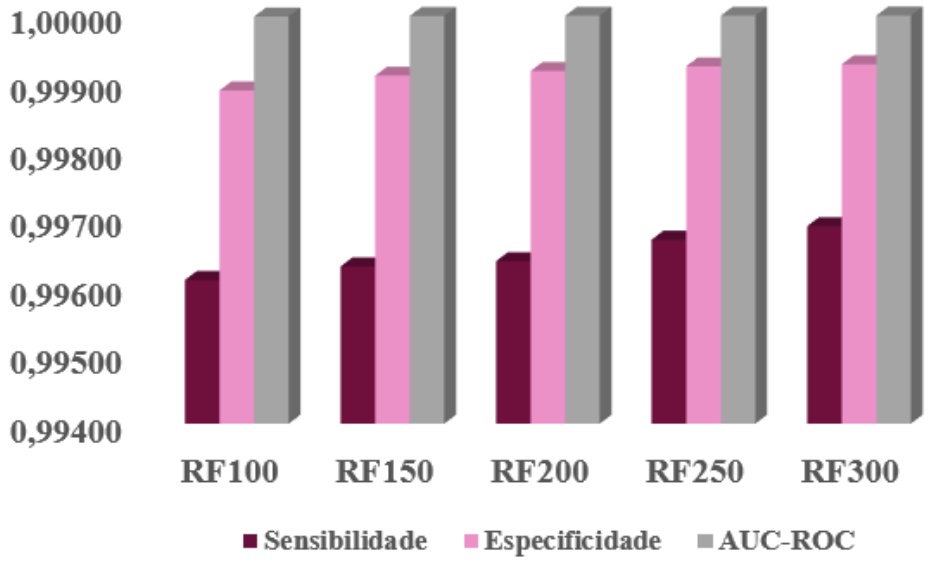

Fonte: O autor (2024)

Gráfico 8 – Índice kappa de treinamento do Random Forest (n=100 até n= 300) para genes acima de 50% de relevância




Fonte: O autor (2024)

Gráfico 9 – Sensibilidade, especificidade e AUC-ROC de treinamento do Random Forest (n=100 até n= 300) para genes acima de 50% de relevância




Fonte: O autor (2024)

### 4.5.2 Teste

A partir da análise do desempenho dos classificadores no treinamento, foram escolhidos o *Random Forest* (n=250) para genes acima de 30% e 40% de relevância, além do *Random Forest* (n=300) para genes acima de 0%, 10%, 40% e 50% de relevância. As duas configurações, com as bases de teste, obtiveram 100% de acurácia, além dos outros indicadores como o índice kappa, sensibilidade, especificidade e AUC-ROC no valor 1, significando que os dois algoritmos conseguiram identificar todos os tipos de câncer corretamente. A matriz de confusão da base de dados de teste pode ser visualizada na Tabela 13.

Tabela 13 – Matriz de confusão da base de dados de teste

| a | b | c | d | e | CLASSES |
|---|---|---|---|---|---|
| 22 | 0 | 0 | 0 | 0 | a = READ |
| 0 | 121 | 0 | 0 | 0 | b = LUAD |
| 0 | 0 | 120 | 0 | 0 | c = THCA |
| 0 | 0 | 0 | 69 | 0 | d = COAD |
| 0 | 0 | 0 | 0 | 255 | e = BRCA |

Fonte: O autor (2024)

A partir do desempenho dos algoritmos nas bases de dados de teste, é possível compreender que com configurações de algoritmos de aprendizado de máquina eficientes, é possível utilizar de genes quantificados, em amostras, para identificar tipos de câncer em diferentes tecidos do corpo feminino. Ademais, essa ferramenta apresenta um potencial auxílio para a área de saúde pública brasileira, como em casos de mulheres que não são recomendadas a fazer mamografia para diagnóstico de câncer de mama, já que com menos de 40 anos, os tecidos mamários são mais densos, resultando em testes de triagem menos confiáveis. A partir da extração de biomarcadores moleculares e as quantificações, seria possível estimar um possível desenvolvimento de câncer de mama a partir de bases de dados previamente treinadas, como as desenvolvidas nesse trabalho.

## 4.6 RELEVÂNCIA DOS GENES PARA O CÂNCER DE MAMA

Logo após os treinamentos e testes com as subamostras, foram feitas mais validações na amostra original com 3 configurações do *Random Forest*, já citadas na seção 4.8, além da primeira (n=10) utilizada para definir as subamostragens, para estudar a relevância dos genes BRCA1 e BRCA2 com interações maiores. A partir disso, foi verificado que até 50 interações, o BRCA1 continuou com relevância nula, enquanto o BRCA2 aumentou a relevância em 5%, mas duplicando o número de interações, o BRCA1 aumentou em 72% a relevância na validação com o BRCA2 constante em 86%, como pode ser visualizado na Tabela 14.

Tabela 14 – Relevância dos genes BRCA1 e BRC2

| CLASSIFICADOR | BRCA1 | BRCA2 |
|---|---|---|
| RF (n = 10) | 0% | 81% |
| RF (n = 20) | 0% | 86% |
| RF (n = 50) | 0% | 86% |
| RF (n = 100) | 72% | 86% |

Fonte: O autor (2024)

# 5 CONCLUSÃO

## 5.1 DESEMPENHO DOS CLASSIFICADORES DE TREINO E TESTE

Durante a etapa de treinamento, foi possível verificar que todos os classificadores e suas configurações específicas obtiveram ótimos resultados, com diferenças ínfimas em valores de centésimos e milésimos. Apesar do melhor desempenho ter sido, no geral, do *Random Forest*, os outros classificadores desempenharam métricas acima de 90%, sendo considerados, também, possíveis classificadores para diferenciação de câncer em mulheres.

Com relação à etapa de teste, o resultado obtido foi bastante satisfatório e esperado com 250 e 300 interações no *Random Forest*, já que essas duas configurações chegaram a quase 100% de acurácia no treinamento, além de sensibilidade e especificidade altas.

## 5.2 IMPORTÂNCIA DOS GENES BRCA1 E BRCA2 PARA O CÂNCER DE MAMA

Os genes BRCA1 e BRCA2, como já descritos no capítulo 2, são reparadores de DNA com funções bastante semelhantes. Em contrapartida, quando verificada a relevância dos genes para a validação, o BRCA1 só obteve um valor significativo a partir de 50 interações, diferentemente do BRCA2 que já apresenta significância com interações menores. A partir disso, pode-se suspeitar que, apesar da mesma função biológica, o BRCA1 e o BRCA2 possam ter importâncias diferentes para o câncer de mama. A partir disso, é possível discutir sobre duas etapas distintas com relação à essa doença: o rastreamento e o diagnóstico. O rastreamento do câncer de mama é feito quando é necessário identificar previamente a doença em mulheres assintomáticas, ou seja, que não apresentem características usuais no aparecimento do câncer. Por outro lado, o diagnóstico é feito quando a paciente já apresenta sintomas e características comuns em mulheres com desenvolvimento do câncer. De acordo com esse contexto, o BRCA1, por não expressar relevância na classificação de interações menores, pode ser mais utilizado para rastreamento da doença, já que com as amostras extraídas de tecidos cancerosos, sua importância para classificação foi nula inicialmente. Ademais, o BRCA2 pode ser hipotetizado como biomarcador de diagnóstico, apresentando mais importância para amostras com características propícias de câncer de mama. Por fim, com o auxílio da Engenharia Genética e Bioinformática, é possível consultar essas informações e aprimorar esse trabalho com dados mais específicos de mulheres com câncer de

mama para que os estudos dos genes BRCA1 e BRCA2 ofereçam resultados ainda mais contundentes e importantes para a saúde feminina.

## 5.3 LIMITAÇÕES DO TRABALHO

A base de dados disponibilizada pelo *Broad Institute* não apresenta informações detalhadas sobre as amostras de tecidos cancerosos, como por exemplo o sexo (masculino e feminino) do paciente, o tempo de estágio do câncer, se há amostras de câncer hereditário e, também, se houve coletas de pacientes inicialmente assintomáticos. Essas especificações podem ser importantes para definir subamostras diferentes, além de auxiliar no estudo da relevância dos genes BRCA1 e BRCA2 com relação ao rastreamento e diagnóstico do câncer de mama. Além disso, os testes de validação para estudo dos genes BRCA1 e BRCA2 foram feitos com a base de dados total e não só com as amostras de câncer de mama. Apesar dos dois biomarcadores moleculares supracitados não serem comumente expressos nos outros tipos de câncer da base de dados, há a possibilidade da presença dos outros 4 cânceres influenciarem os resultados específicos das relevâncias dos genes nas configurações propostas no *Random Forest*.

# REFERÊNCIAS

ABDULLAH, M. I. et al. Papillary Thyroid Cancer: Genetic Alterations and Molecular Biomarker Investigations. **International Journal of Medical Sciences**, v. 16, n. 3, p. 450–460, 28 fev. 2019.

ALBERTS, B. et al. **Biologia Molecular da Célula**. [s.l.] Artmed Editora, 2017.

BARZAMAN, K. et al. Breast cancer: Biology, biomarkers, and treatments. **International Immunopharmacology**, v. 84, n. 106535, p. 106535, jul. 2020.

**Broad GDAC Firehose**. Disponível em: <https://gdac.broadinstitute.org>. Acesso em: 6 Jul. 2024.

Câncer - OPAS/OMS**. Organização Pan-Americana da Saúde**. 2020. Disponível em: <https://www.paho.org/pt/topicos/cancer#:~:text=O%20c>. Acesso em: 19 Jun. 2024

Carga global de câncer aumenta em meio à crescente necessidade de serviços - OPAS/OMS. **Organização Pan-Americana da Saúde**. 2024. Disponível em: <https://www.paho.org/pt/noticias/1-2-2024-carga-global-cancer-aumenta-em-meio-crescente-necessidade-servicos>. Acesso em: 19 Jun. 2024.

COHEN, E. R. et al. CD44 and associated markers in oral rinses and tissues from oral and oropharyngeal cancer patients. **Oral Oncology**, v. 106, p. 104720, 1 jul. 2020.

Como os diferentes tipos de câncer são originados?. **Pfizer Brasil**. 2022. Disponível em: <https://www.pfizer.com.br/noticias/ultimas-noticias/como-os-tipos-de-cancer-sao-originados>.

Como se comportam as células cancerosas?. **Instituto Nacional do Câncer**. Disponível em: <https://www.gov.br/inca/pt-br/assuntos/cancer/como-surge-o-cancer/como-se-comportam-as-celulas-cancerosas>.

DEMARCO, C. **Adenocarcinomas: 6 things to know about the "cancer of the cavities"**. Disponível em: <https://www.mdanderson.org/cancerwise/adenocarcinomas--6-things-to-know-about-the--cancer-of-the-cavities.h00-159465579.html>. Acesso em: 19 Jun. 2024.

Em 2021, 62% dos pacientes com câncer tiveram diagnóstico tardio no país. **Instituto Oncoguia**. Disponível em: <https://www.oncoguia.org.br/conteudo/em-2021-62-dos-pacientes-com-cancer-tiveram-diagnostico-tardio-no-pais/16465/7/#:~:text=Em%202008%2C%2053%25%20dos%20pacientes>.Acesso em: 19 jul. 2024.

Entenda os tipos de câncer de pele. **G1**. Disponível em: <https://g1.globo.com/bemestar/noticia/2019/12/07/entenda-os-tipos-de-cancer-de-pele.ghtml>. Acesso em: 19 Jun. 2024.

Estimativa 2023: incidência de câncer no Brasil. **Instituto Nacional de Câncer**. 2023. Disponível em:

<https://www.inca.gov.br/publicacoes/livros/estimativa-2023-incidencia-de-cancer-no-brasil>. Acesso em: 19 Jun. 2024.

FARIA, Á. **Alterações genéticas: Glossário de genética III**. Disponível em: <https://blog.mendelics.com.br/glossario-de-genetica-parte-3/>. Acesso em: 19 Jun. 2024.

FREEDMAN, D. A. **Statistical Models**. [S.l.: s.n.], 2009. ISBN 9781139477314.

HALLE, B. R.; JOHNSON, D. B. Defining and Targeting BRAF Mutations in Solid Tumors. **Current Treatment Options in Oncology**, v. 22, n. 4, 27 fev. 2021.

HAN, S.-A.; KIM, S.-W. BRCA and Breast Cancer-Related High-Penetrance Genes. **Advances in Experimental Medicine and Biology**, v. 1187, p. 473–490, 2021.

HELM, J. M. et al. Machine Learning and Artificial Intelligence: Definitions, Applications, and Future Directions. **Current Reviews in Musculoskeletal Medicine**, v. 13, n. 1, p. 69–76, 25 jan. 2020.

JIANG, T.; GRADUS, J. L.; ROSELLINI, A. J. Supervised Machine Learning: A Brief Primer. **Behavior Therapy**, v. 51, n. 5, p. 675–687, set. 2020.

KHAN, H. et al. Cancer biomarkers and their biosensors: A comprehensive review. **TrAC Trends in Analytical Chemistry**, p. 116813, nov. 2022.

KIM, E.-K.; SHIN, H.-C. The Potential Predictors in Chemotherapy Sensitivity. **Advances in experimental medicine and biology**, p. 381–389, 1 jan. 2021.

KONTOMANOLIS, E. N. et al. Role of Oncogenes and Tumor-suppressor Genes in Carcinogenesis: a Review. **Anticancer Research**, v. 40, n. 11, p. 6009–6015, 1 nov. 2020.

LIN, Y.; XU, J.; LAN, H. Tumor-associated macrophages in tumor metastasis: biological roles and clinical therapeutic applications. **Journal of Hematology & Oncology**, v. 12, n. 1, 12 jul. 2019.

MALKI, A. et al. Molecular Mechanisms of Colon Cancer Progression and Metastasis: Recent Insights and Advancements. **International Journal of Molecular Sciences**, v. 22, n. 1, 24 dez. 2020.

MCHUGH, M. L. Interrater reliability: the kappa statistic. **Biochemia Medica, Croatian Society for Medical Biochemistry and Laboratory Medicine**. p. 276–282, 2012.

MOORTHY, C. G.; SANKAR, G. U. Numerical methods for calculus students. **[S.l.]: LAP Lambert Academic Publishing**, 2019.

MORI, Y. et al. Deep learning-based gene selection in comprehensive gene analysis in pancreatic cancer. **Scientific Reports**, v. 11, n. 1, 13 ago. 2021.

MOUBARZ, G. et al. Lung cancer risk in workers occupationally exposed to polycyclic aromatic hydrocarbons with emphasis on the role of DNA repair gene. **International Archives of Occupational and Environmental Health**, v. 96, n. 2, p. 313–329, 26 out. 2022.

NITHYA, P.; CHANDRASEKAR, A. NBN Gene Analysis and it's Impact on Breast Cancer. **Journal of Medical Systems**, v. 43, n. 8, 5 jul. 2019.

OLIVEIRA, K. C. S. et al. Current Perspectives on Circulating Tumor DNA, Precision Medicine, and Personalized Clinical Management of Cancer. **Molecular Cancer Research**, v. 18, n. 4, p. 517–528, 1 abr. 2020.

O que é aprendizado por reforço? — Explicação do aprendizado por reforço. **Amazon Web Services**. Disponível em: <https://aws.amazon.com/pt/what-is/reinforcement-learning/#:~:text=O%20aprendizado%20por%20refor>. Acesso em: 18 Jun. 2024.

OTMANI, K.; LEWALLE, P. Tumor Suppressor miRNA in Cancer Cells and the Tumor Microenvironment: Mechanism of Deregulation and Clinical Implications. **Frontiers in Oncology**, v. 11, 15 out. 2021.

PATEL, A. Benign vs Malignant Tumors. **JAMA Oncology**, v. 6, n. 9, 30 jul. 2020.

Percepções e prioridades do câncer nas favelas brasileiras. **Instituto Oncoguia**. 2023. Disponível em: <https://www.oncoguia.org.br/conteudo/percepcoes-e-prioridades-do-cancer-nas-favelas-brasileiras/16284/1093/>. Acesso em: 20 Jun. 2024.

ROY, S. et al. Classification models for Invasive Ductal Carcinoma Progression, based on gene expression data-trained supervised machine learning. **Scientific Reports**, v. 10, n. 1, p. 4113, 5 mar. 2020.

SEARS, C. R.; MAZZONE, P. J. Biomarkers in Lung Cancer. **Clinics in Chest Medicine**, v. 41, n. 1, p. 115–127, mar. 2020.

Tumor Benigno: Entenda O Que É e Os Sintomas Da Neoplasia Benigna | Vida Saudável |. **Hospital Israelita Albert Einstein**. Disponível em: https://vidasaudavel.einstein.br/tumor-benigno-entenda-o-que-e-os-sintomas-da-neoplasia-benigna/#:~:text=Exemplos comuns de tumores benignos. Acesso em: 14 maio. 2024.

WANG, J. et al. MI_DenseNetCAM: A Novel Pan-Cancer Classification and Prediction Method Based on Mutual Information and Deep Learning Model. **Frontiers in Genetics**, v. 12, 3 jun. 2021.

WU, J.; HICKS, C. Breast Cancer Type Classification Using Machine Learning. **Journal of Personalized Medicine**, v. 11, n. 2, p. 61, 20 jan. 2021.