# Sistema de Predição do Diagnóstico de Câncer de Mama Utilizando o Classificador Ingênuo de Bayes


1st Alexandre Burle
*Centro de Informática*
*Universidade Federal de Pernambuco*
Recife, Brasil
aqb@cin.ufpe.br

2nd Lucas Morais
*Centro de Informática*
*Universidade Federal de Pernambuco*
Recife, Brasil
lma6@cin.ufpe.br

3rd Marco Aurélio
*Centro de Informática*
*Universidade Federal de Pernambuco*
Recife, Brasil
mafs3@cin.ufpe.br

4th Matheus Andrade
*Centro de Informática*
*Universidade Federal de Pernambuco*
Recife, Brasil
mvtna@cin.ufpe.br


## I. Justificativa

O câncer de mama é responsável por 1 em cada 4 casos de câncer entre mulheres no mundo. Em 2020, foram registrados cerca de 2,2 milhões de novos casos de câncer de mama ao redor do mundo para a população feminina (Figura 1). Neste mesmo período, a estimativa de mortes por essa doença atingiu a marca de 684.996 mulheres (Figura 2), revelando-se como uma das principais causas de morte por câncer em pessoas do sexo feminino[1, 2].

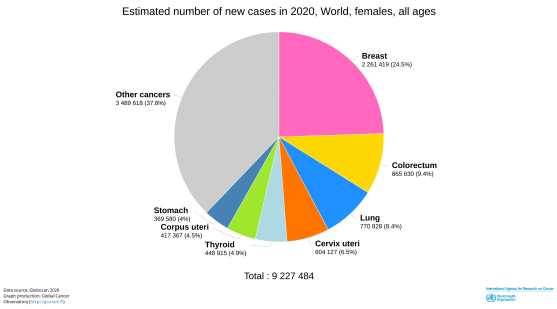


Fig. 1. Novos casos de câncer em mulheres no mundo em 2020.

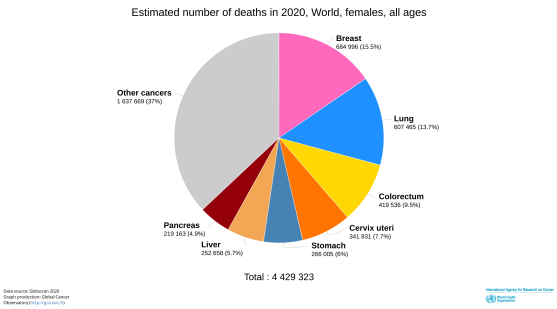


Fig. 2. Mortes de mulheres por câncer em 2020.

A presença de sinais anormais na mama, tais como nódulos mamários, retração na pele da mama, alteração no formato do mamilo ou aumento progressivo do tamanho da mama com a presença de sinais de edema, torna essencial a investigação do órgão para se determinar o diagnóstico do paciente. O diagnóstico precoce do câncer de mama é fundamental para prevenir a progressão da doença para estágios avançados e reduzir a taxa de mortalidade associada a essa doença.[3]

## II. Objetivos

As abordagens utilizadas para a detecção do câncer de mama são o diagnóstico precoce e o rastreamento. O diagnóstico precoce é utilizado em pessoas com sinais e/ou sintomas iniciais da doença. O rastreamento é uma estratégia direcionada a mulheres em uma faixa etária específica e com uma periodicidade determinada, as quais não apresentam sinais indicativos da doença[4]. A existência de rastreamento, mesmo com boa cobertura, não prescinde das estratégias de diagnóstico precoce, pois são abordagens complementares[1].

Para a investigação, são utilizados exames clínicos das mamas e exames de imagem como mamografia, ultrassonografia ou ressonância magnética. No entanto, a confirmação diagnóstica só é obtida através de biópsia, onde ocorre a retirada - através de punção ou cirurgia - e a análise de uma porção do corpo suspeito por parte de um patologista.

A detecção precoce do câncer de mama é crucial para o sucesso do tratamento, assim como para a sobrevivência dos pacientes[4] e é diretamente relacionada com a precisão do diagnóstico da doença, uma vez que a eficiência dos resultados clínicos contribuem para redução da mortalidade. Assim, obter o diagnóstico do câncer de maneira efetiva é uma prioridade para profissionais de saúde. Nesse cenário, os sistemas de previsão de diagnóstico de câncer de mama têm se mostrado ferramentas importantes para auxiliar na identificação precoce e no tratamento da doença, uma vez que são capazes de indicar diagnósticos a partir de informações do paciente e dos nódulos suspeitos[5].

O objetivo deste documento é, portanto, apresentar o desenvolvimento e a análise de um sistema de previsão de diagnóstico de câncer de mama baseado em um Classificador Ingênuo de Bayes utilizando o conjunto de dados *Breast Cancer Wisconsin (Diagnostic) Data Set*, criado por pesquisadores da Universidade de Wisconsin e distribuído no repositório de aprendizagem de máquina da Universidade da Califórnia em Irvine (UCI)[6]. O modelo proposto neste documento será desenvolvido a partir do *Gaussian Naive Bayes* fornecido pela biblioteca de código *scikit-learn*[7], disponível para linguagem de programação Python[8], de acordo com as métricas de Acurácia, Precisão, Sensibilidade (*Recall*) e F1-Score, cujas formulações são definidas pelas equações 1, 2, 3 e 4, respectivamente, definidas na seção III-D.

## III. Metodologia

### A. Dataset

O treinamento e teste do sistema desenvolvido nesse documento é realizado com base no *dataset Breast Cancer Wisconsin (Diagnostic) Data Set*, criado por pesquisadores da Universidade de Wisconsin e distribuído no repositório de aprendizagem de máquina da UCI. O *dataset* é formado pelas características de imagens digitalizas de amostras de tecido da mama obtidos por meio de aspiração por agulha fina (PAAF), um procedimento utilizado para se obter amostra de células, tecido ou fluidos de uma área do corpo utilizando uma agulha. Com a amostra das células da mama obtidas, é utilizada uma árvore de decisão ("*Multisurface Method-Tree*" (MSM-T) [9]) para fazer a separação da célula em planos, e um outro programa, descrito por K. P. Bennett e O. L. Mangasarian[10], foi utilizado para obter os planos de separação num espaço tridimensional.

No total, o conjunto de dados contém 569 amostras, cada uma formada por 32 atributos (que em todas as amostras possuem valores válidos e não nulos). A descrição dos atributos de cada amostra é apresentada na Tabela I, na qual os atributos de c) a l) são valores reais e se repetem para cada um dos três núcleos celulares.

Tabela I
Informações dos Atributos.

| Índice | Nome | Descrição | Tipo |
|---|---|---|---|
| a) | Número de Identificação | ID | Inteiro |
| b) | Diagnóstico | Indica o resultado da doença | Categórica: M = maligno, B = benígno |
| c) | Raio | Média das distâncias do centro aos pontos do perímetro | Número Real |
| d) | Textura | Desvio padrão dos valores da escala de cinza | Número Real |
| e) | Perímetro | Perímetro do núcleo celular | Número Real |
| f) | Área | Área do núcleo celular | Número Real |
| g) | Suavidade | Variação local em comprimentos de raio | Número Real |
| h) | Compacidade | $\left(\frac{perimetro^2}{area-1.0}\right)$ | Número Real |
| i) | Concavidade | Gravidade das porções côncavas do contorno | Número Real |
| j) | Pontos côncavos | Número de porções côncavas do contorno | Número Real |
| k) | Simetria | Grau de simetria do núcleo celular | Número Real |
| l) | Dimensão fractal | ("*coastline approximation*" - 1) | Número Real |

### B. Classificador Ingênuo de Bayes

O Classificador Ingênuo de Bayes (*Naive* Bayes)[11] é um algoritmo de aprendizagem de máquina que utiliza a probabilidade para classificar dados em diferentes categorias. Ele é conhecido por ser um dos algoritmos mais simples e eficazes para a classificação de dados.

O *Naive* Bayes é um dos modelos mais conhecidos a aplicar o conceito de probabilidade e utiliza o Teorema de Bayes[11] como princípio fundamental, sendo definido como:

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

Sendo:

- $A, B$: Dois eventos contidos no espaço amostral $S$ de um experimento $\epsilon$;
- $P(A|B)$: A probabilidade condicional de A dado B;
- $P(B|A)$: A probabilidade condicional de B dado A;
- $P(A)$: A probabilidade marginal de A;
- $P(B)$: A probabilidade marginal de B.

O algoritmo parte da premissa de que as características (ou atributos) dos dados são independentes entre si. Essa independência é considerada "ingênua", visto que na realidade muitas características estão interconectadas.

O Classificador Ingênuo de Bayes funciona calculando a probabilidade de cada categoria com base nos valores das características do dado a ser classificado. Ele então seleciona a categoria com a maior probabilidade como a categoria para o dado.

### C. Tratamento dos Dados

Considerando que o *dataset* utilizado não possui campos nulos ou inválidos, o único atributo que será desconsiderado é o número de identificação da amostra, pois não fornece nenhuma informação útil para a classificação da amostra. Ademais, os valores das amostras serão utilizados da forma que foram fornecidos no site da Universidade de Winsconsin.

As 569 amostras do *dataset* tem a seguinte distribuição entre duas classes: 357 benígnas e 212 malígnas. Sendo assim, 80% de cada classe será selecionada aleatoriamente para treinamento e os 20% restantes para teste.

### D. Métricas

As métricas que fornecem informações para a análise da performance do modelo serão baseadas na Matriz de Confusão

do sistema. Uma Matriz de Confusão apresenta quatro fatores referentes às predições do modelo para com o *dataset* de testes. A estrutura de uma Matriz de Confusão é apresentada na Tabela II.

Tabela II
MATRIZ DE CONFUSÃO

| | | Classe Prevista | |
|---|---|---|---|
| | | Positivo | Negativo |
| Classe Real | | TP | FN |
| | | FP | TN |

Sendo:

- **Verdadeiro Positivo (TP)**: Representam as predições em que o modelo indicou corretamente um diagnóstico positivo para câncer;
- **Verdadeiro Negativo (TN)**: Representam as predições que o modelo indicou corretamente um diagnóstico negativo para câncer;
- **Falso Positivo (FP)**: Representam as predições em que o modelo indicou incorretamente um diagnóstico positivo de câncer;
- **Falso Negativo (FN)**: Representam as predições em que o modelo indicou incorretamente um diagnóstico negativo de câncer.

As métricas utilizadas para avaliar o desempenho do classificador são Acurácia, Recall ou Sensibilidade, Especificidade e Precisão.

*1) Acurácia:* A acurácia fornece uma informação do desempenho geral do modelo. Essa métrica mapeia a performance do sistema a partir da razão entre as predições corretas do sistema e as predições totais. Portanto, a acurácia do sistema é determinada a partir da seguinte expressão

$$Acurácia = \frac{TP + TN}{TP + TF + FP + FN} \tag{1}$$

*2) Precisão:* A previsão fornece uma informação referente à quantidade de predições positivas que são, de fato, positivas. Essa métrica também é conhecida como Valor Preditivo Positivo (VPP). Portanto, a precisão do modelo é representada a partir da seguinte expressão

$$Precisão = \frac{TP}{TP + FP} \tag{2}$$

*3) Recall ou Sensibilidade:* A sensibilidade fornece uma informação referente à capacidade do sistema em detectar com sucesso resultados classificados como positivos. Dessa forma, a sensibilidade do modelo é expressa a partir da seguinte expressão

$$Sensibilidade = \frac{TP}{TP + FN} \tag{3}$$

*4) F1-Score:* O F1-Score fornece uma informação referente a uma média entre as métricas de sensibilidade e precisão. Dessa forma, o F1-Score do modelo é expressa a partir da seguinte expressão

$$F1 - Score = \frac{2TP}{2TP + FP + FN} \tag{4}$$

### *E. Desenvolvimento do Sistema*

A implementação do modelo proposto será em Python utilizando o ambiente de desenvolvimento do Google Colaboratory[12]. Para o tratamento inicial e manipulações dos dados será utilizada a biblioteca Pandas[13].

*1) Pré-Processamento dos Dados:* Os dados originais do *dataset* são devidamente preparados, sendo ajustados e transformados para formatos úteis de manusear durante o treinamento do modelo, configurando, assim, a etapa de pré-processamento dos dados.

*2) Análise Exploratória dos Dados:* A fase inicial do desenvolvimento do modelo se baseia na análise exploratória dos dados. É interessante visualizar e compreender como os dados estão estruturados de modo que possibilite a análise de estratégias para manipulação e experimentos a serem feitos com o conjunto de dados.

*3) Outliers:* Ao visualizar as informações da tabela, percebe-se que várias colunas possuem dados com valores que destoam muito dos demais e podem causar anomalias nos resultados obtidos, são os chamados *outliers*. Para resolver esse problema, todas as linhas da tabela que contém atributos a uma distância superior a 3 desvios padrão da média foram removidas. Um exemplo de gráfico boxplot, o qual contém uma visualização dos *outliers* do atributo raio dos três núcleos celulares, é apresentada na Figura 3.

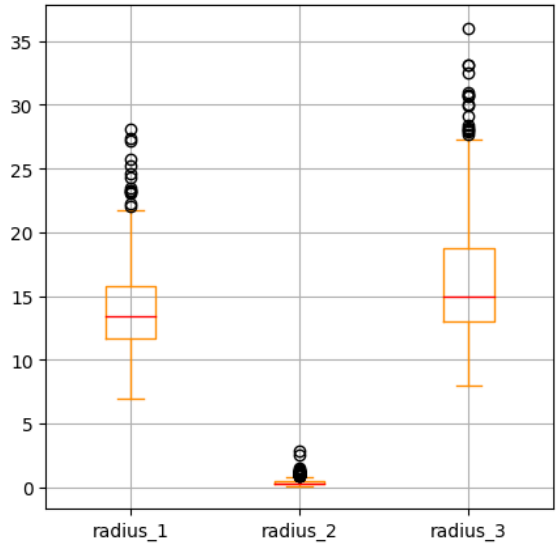


Fig. 3. Exemplo de boxplot do atributo raio de cada núcleo celular.

*4) Análise Gráfica e Estatística:* A fim de visualizar melhor as distribuições dos dados, utiliza-se uma representação gráfica das amostras, as quais revelam a distribuição dos dados totais presentes no *dataset*. Os dados numéricos (todos os atributos, exceto *ID* e *Compacidade*) foram representados por gráficos de barras.

Além disso, é realizada uma análise estatística para facilitar a compreensão das características dos pacientes com diagnóstico positivo para câncer de mama, auxiliando no

entendimento da relação entre os atributos dos dados e o seu diagnóstico.Para isso, aplica-se um filtro no *dataset* para retornar apenas os pacientes com diagnóstico positivo e, munidos deste subconjunto, realiza-se a distribuição dos dados de acordo com cada característica, além de identificar a média dessas categorias.

Exemplos de representações gráficas referentes à análise de distribuição dos dados do *dataset*, e à análise da relação entre as características dos pacientes com diagnóstico positivo para o câncer de mama e este resultado são apresentados nas Figuras 4 e 5, respectivamente.

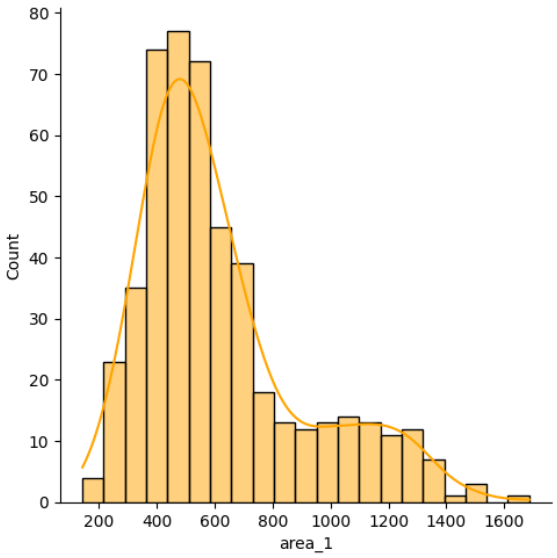


Fig. 4. Exemplo de representação gráfica da distribuição do atributo de área do primeiro núcleo celular.

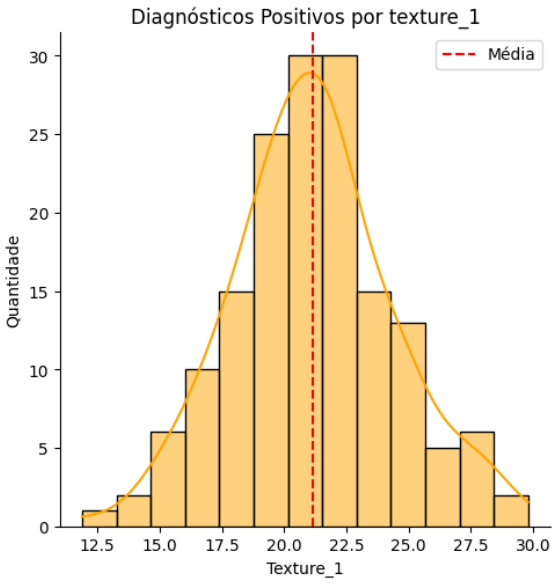


Fig. 5. Exemplo de representação gráfica da distribuição do atributo de textura do primeiro núcleo celular para pacientes com diagnóstico positivo.

*5) Treinamento:* Para a etapa de treinamento, é necessário separar o *dataset* em dois grupos de dados, sendo o primeiro responsável pelo aprendizado do modelo (*dataset* de treino), e o segundo por realizar os testes (*dataset* de teste), assim como descrito na seção III-A.

A interface de treinamento definida permite a estratificação dos dados no treino. Esse fator é utilizado para os diferentes cenários de teste que será apresentado na seção IV.

Além disso, durante o treinamento do modelo é utilizada a validação cruzada. A validação cruzada é uma técnica importante para avaliar o desempenho do modelo em dados não vistos, sem comprometer um conjunto de dados que deve ser mantido para testes finais. Essa técnica envolve a divisão dos dados em conjuntos de treinamento e validação, onde o modelo é ajustado com o conjunto de treinamento e avaliado com o conjunto de validação. Esse processo é repetido várias vezes, cada vez com um conjunto de treinamento e validação diferentes, e o desempenho do modelo é calculado a partir da média dos resultados obtidos em cada iteração. A validação cruzada é útil para evitar o *overfitting*. Para este modelo, utiliza-se uma quantidade de divisões do *dataset* em treino e validação de 5 iterações.

## IV. Experimentos

Para a análise da performance do modelo, realiza-se diferentes cenários de treino e teste. As abordagens experimentais utilizadas seguem uma estrutura de cenários básicos e cenários adicionais. Os cenários adicionais se baseiam nas abordagens básicas e se fundem para resultar em possibilidades mais complexas.

### A. Cenários Básicos

As hipóteses básicas de teste do modelo são: *dataset* bruto, remoção do atributo *ID*, remoção do atributo de compacidade e seleção de características usando *Random Forest*[14]. Cada um desses conjuntos de testes citados serão descritos a seguir.

*1) Conjunto de Dados Brutos:* O primeiro experimento é baseado no treino e teste da performance do algoritmo para a base de dados após a remoção dos *outliers*, sem modificação. Espera-se que o resultado desse experimento sirva como *baseline* para os demais testes, visto que representa a situação mais simples a ser realizada.

*2) Remoção do Atributo ID:* O segundo experimento está relacionado com a remoção do atributo *ID* do conjunto de dados. Este atributo é único para cada amostra do *dataset*. Dessa forma, percebe-se que este valor não fornece uma informação interessante para a aprendizagem do modelo.

*3) Remoção do Atributo de Compacidade:* O terceiro experimento é definido pela remoção do atributo de compacidade da base de dados, além da remoção do *ID* realizada anteriormente. O atributo de compacidade é definido como $\left(\frac{perimetro^2}{area-1.0}\right)$.

Dessa forma, percebe-se que a compacidade depende da área e do perímetro, o que não é interessante para o algoritmo Naive Bayes, visto que a suposição do Classificador Ingênuo de Bayes é que as características são independentes.

*4) Seleção de Características Utilizando Random Forest:* O quarto cenário de teste se configura pelo uso de uma *Random Forest*[14] para seleção das características do conjunto de dados que são mais relevantes para o treinamento do modelo.

A *Random Forest* constrói várias árvores de decisão com diferentes subconjuntos de *features* aleatórios, e as *features* que são mais importantes para a predição do modelo são selecionadas com base na sua frequência de aparecimento nas árvores. A seleção de *features* é importante para melhorar a precisão do modelo, reduzir o tempo de treinamento e evitar o *overfitting*.

### B. Cenários Adicionais

O conjunto de dados resultante após a remoção dos *outliers* enfatizou o desbalanceamento entre as classes de diagnósticos positivo (Maligno - M) e negativo (Benigno - B) para o câncer de mama, revelando que a quantidade de amostras com diagnóstico positivo para câncer de mama (160) representa menos da metade da quantidade de amostras com diagnóstico negativo (327).

Dessa forma, utiliza-se a estratégia de estratificação dos dados no treinamento e teste para cada uma das abordagens básicas apresentadas em IV-A. Além disso, a partir dos cenários resultantes (dados não-estratificados e dados estratificados) é testado um cenário adicional: o balanceamento do *dataset*.

*1) Estratificação dos Dados:* A estratificação dos dados divide o conjunto de dados de modo que se mantenha a proporção entre as classes presentes no conjunto original. Essa separação é realizada de acordo com uma coluna do *dataset*. No caso deste modelo, a coluna escolhida é a coluna de diagnóstico, a qual representa a predição do modelo. A importância da estratificação se revela, principalmente, quando os dados estão muito desbalanceados, que é o caso do conjunto de dados utilizado.

*2) Balanceamento dos Dados:* O balanceamento do *dataset* é o experimento final deste documento. Esse cenário baseia-se na utilização do algoritmo *Random Oversampling*[15] para balancear o conjunto de dados.

## V. Resultados

Os resultados dos experimentos realizados são analisados de acordo com as métricas de Acurácia, Precisão, Sensibilidade e F1-Score, apresentadas em III-D. Os resultados referentes aos cenários básicos com dados desbalanceados e balanceados são apresentados nas Tabelas III e IV, respectivamente.

Analisando o processo de treinamento do modelo e, especialmente, investigando os resultados dos experimentos realizados, conclui-se pontos interessantes sobre a performance e sobre as hipóteses experimentais.

Utilizando como base as métricas definidas no início deste documento, na seção II, tem-se que o modelo de melhor performance foi o classificador treinado com os dados não balanceados, não estratificados e com a remoção os atributos de ID e Compacidade. Esse classificador apresentou uma porcentagem de 96% para todas as métricas analisadas e um desvio padrão de 2%.

Tabela III
Resultados dos Experimentos com Dados Desbalanceados, Estratificados (E) e Não-Estratificados (NE).

| Cenário | Acurácia Média | Desvio Padrão | Acurácia | Precisão | Recall | F1-Score |
|---|---|---|---|---|---|---|
| Dados Brutos (NE) | 0.73 | 0.03 | 0.71 | 0.81 | 0.62 | 0.60 |
| Dados Brutos (E) | 0.73 | 0.06 | 0.78 | 0.88 | 0.66 | 0.67 |
| Remoção do ID (NE) | 0.94 | 0.02 | 0.94 | 0.93 | 0.94 | 0.94 |
| Remoção do ID (E) | 0.94 | 0.02 | 0.94 | 0.94 | 0.92 | 0.93 |
| Remoção da Compacidade (NE) | 0.96 | 0.02 | 0.96 | 0.96 | 0.96 | 0.96 |
| Remoção da Compacidade (E) | 0.95 | 0.02 | 0.95 | 0.95 | 0.93 | 0.94 |
| Random Forest (NE) | 0.94 | 0.02 | 0.95 | 0.94 | 0.95 | 0.94 |
| Random Forest (E) | 0.94 | 0.03 | 0.94 | 0.94 | 0.92 | 0.93 |

Tabela IV
Resultados dos Experimentos com Dados Balanceados, Estratificados (E) e Não-Estratificados (NE).

| Cenário | Acurácia Média | Desvio Padrão | Acurácia | Precisão | Recall | F1-Score |
|---|---|---|---|---|---|---|
| Dados Brutos (NE) | 0.81 | 0.06 | 0.79 | 0.85 | 0.79 | 0.78 |
| Dados Brutos (E) | 0.84 | 0.03 | 0.82 | 0.85 | 0.82 | 0.82 |
| Remoção do ID (NE) | 0.93 | 0.04 | 0.92 | 0.93 | 0.93 | 0.92 |
| Remoção do ID (E) | 0.93 | 0.01 | 0.94 | 0.94 | 0.94 | 0.94 |
| Remoção da Compacidade (NE) | 0.94 | 0.03 | 0.94 | 0.94 | 0.94 | 0.94 |
| Remoção da Compacidade (E) | 0.94 | 0.01 | 0.92 | 0.93 | 0.92 | 0.92 |
| Random Forest (NE) | 0.81 | 0.06 | 0.79 | 0.85 | 0.79 | 0.78 |
| Random Forest (E) | 0.84 | 0.03 | 0.82 | 0.85 | 0.82 | 0.82 |

## VI. Conclusão

Com base nos resultados apresentados neste estudo, observa-se que o cenário que apresentou o modelo de melhor desempenho (classificador treinado com os dados não balanceados, não estratificados e com a remoção os atributos de ID e Compacidade) foi um dos cenários onde a estratificação dos dados não melhorou o resultado do experimento seguindo a mesma abordagem básica. Além disso, a hipótese envolvendo a seleção de características em dados não balanceados, assim como o experimento baseado na remoção dos atributos de ID e Compacidade em dados balanceados também levaram a uma redução na performance quando a estratificação dos dados foi aplicada. Portanto, pode-se concluir que a estratificação dos dados não é uma garantia de melhora na performance do modelo, apesar de ser uma estratégia frequentemente recomendada para conjuntos de dados desbalanceados.

Além disso, o balanceamento do conjunto de dados não forneceu melhorias significativas para a performance do modelo. Para conjuntos de dados com uma pequena diferença numérica entre as quantidades de amostras de cada classe, uma abordagem interessante seria a remoção aleatória de dados da classe majoritária, em vez da criação aleatória de dados da classe minoritária. Essa estratégia é interessante, pois pode evitar a introdução de ruído no conjunto de dados balanceado.

## VII. Cronograma de Atividades

As atividades relacionadas à pesquisa, desenvolvimento, análise e apresentação do sistema proposto serão divididas de acordo com os pilares necessários para o andamento do projeto. O cronograma das tarefas a serem realizadas juntamente com os seus prazos estão apresentadas na Tabela V.

Tabela V
CRONOGRAMA DE ATIVIDADES.

| Período | Descrição da Atividade |
|---|---|
| 15/02 - 20/02 | Pesquisa e escolha do *Dataset*. |
| 21/02 - 27/02 | Decisão do tema. |
| 28/02 - 10/03 | Construção da proposta do projeto. |
| 11/03 - 15/03 | Análise da base de dados e estudo das ferramentas. |
| 16/03 - 30/03 | Construção do modelo proposto usando o Classificador Ingênuo de Bayes. |
| 31/03 - 05/03 | Elaboração e execução de testes. |
| 06/04 - 08/04 | Estudo crítico dos resultados obtidos. |
| 09/03 - 12/04 | Elaboração do relatório de projeto. |
| 13/04 - 16/04 | Slides de apresentação. |

## REFERÊNCIAS


[1] A. Migowski *et. al.* "Diretrizes para detecção precoce do câncer de mama no Brasil. III - Desafios à implementação". In: (2018).

[2] World Health Organization (WHO). *International Agency for Research on Cancer: Breast 2020 Report*. URL: https://gco.iarc.fr/today/data/factsheets/cancers/20-Breast-fact-sheet.pdf (visited on 03/06/2023).

[3] Instituto Nacional de Câncer. *Detecção precoce*. 2022. URL: https://www.gov.br/inca/pt-br/assuntos/gestor-e-profissional-de-saude/controle-do-cancer-de-mama/acoes/deteccao-precoce (visited on 03/06/2023).

[4] World Health Organization (WHO). *Cancer Control Knowledge into Action WHO Guide for Effective Programmes*. Vol. 3. Cancer Control Series. 20 Avenue Appia, 1211 Geneva 27, Switzerland: WHO Press, 2007.

[5] Srwa Hasan, Ali Sagheer, and Hadi Veisi. "Breast Cancer Classification Using Machine Learning Techniques: A Review". In: *Turkish Journal of Computer and Mathematics Education (TURCOMAT)* 12 (Sept. 2021), pp. 1970–1979.

[6] UCI. *Breast Cancer Wisconsin (Diagnostic) Data Set*. URL: https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%5C%28Diagnostic%5C%29 (visited on 03/06/2023).

[7] *Naive Bayes*. URL: https://scikit-learn.org/stable/modules/naive_bayes.html (visited on 03/06/2023).

[8] *Python*. URL: https://www.python.org (visited on 03/06/2023).

[9] Kristin Bennett. "Decision tree construction via linear programming". In: *Proceedings of the 4th Midwest Artificial Intelligence and Cognitive Science Society Conference, Utica, Illinois* (Jan. 1992).

[10] Kristin P. Bennett and O. L. Mangasarian. "Robust linear programming discrimination of two linearly inseparable sets". In: *Optimization Methods and Software* 1.1 (1992), pp. 23–34. DOI: 10.1080/10556789208805504. eprint: https://doi.org/10.1080/10556789208805504. URL: https://doi.org/10.1080/10556789208805504.

[11] S. Ranganathan, K. Nakai, and C. Schonbach. *Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics*. Elsevier Science, 2018, p. 403. ISBN: 9780128114322. URL: https://books.google.com.br/books?id=rs51DwAAQBAJ.

[12] *Google Colaboratory*. URL: https://colab.research.google.com (visited on 03/06/2023).

[13] *Pandas lib*. URL: https://pandas.pydata.org (visited on 03/06/2023).

[14] sklearn. *Random Forest*. URL: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html (visited on 04/20/2023).

[15] sklearn. *Random OverSampler*. URL: https://imbalanced-learn.org/stable/references/generated/imblearn.over_sampling.RandomOverSampler.html (visited on 04/20/2023).