Sistema de Predição de Doenças Cardíacas Utilizando o Classificador Ingênuo de Bayes

Maria Clara Alves Acruchi
Centro de Informática
UFPE
Recife, Brasil
mcaa@cin.ufpe.br

Maria Luísa dos Santos Silva
Centro de Informática
UFPE
Recife, Brasil
mlss@cin.ufpe.br

Vinícius Sales Oliveira
Centro de Informática
UFPE
Recife, Brasil
vso2@cin.ufpe.br

***Resumo*—Este trabalho teve como objetivo a criação de um sistema de predição da incidência de doenças cardíacas utilizando o Classificador Ingênuo de Bayes. São apresentados o algoritmo classificador, a base de dados utilizada e o processo de modelagem do classificador. Por fim, temos a implementação e relatório do sistema de predição.**

***Index Terms*—doenças cardíacas, predição, classificador ingênuo de Bayes, Naïve Bayes**

I. Introdução

As doenças cardiovasculares representam o grupo de doenças que atingem o coração e também os vasos sanguíneos. Segundo a Estimativa Mundial da Saúde, da Organização Mundial da Saúde (OMS) publicada em dezembro de 2019, a doença cardíaca permanece sendo a principal causa de morte ao redor do mundo nos últimos 20 anos. O número de mortes por doenças cardiovasculares não para de crescer e aumentou em 7 milhões, comparado ao censo do ano 2000 [4].

A causa exata das cardiopatias, ou seja, as doenças cardiovasculares, não é clara, mas existem muitas maneiras de aumentar o risco de desenvolver essas patologias. São chamados "fatores de risco". Quanto mais desses fatores uma pessoa tem, maior é a possibilidade de desenvolver doenças relacionadas ao coração [2].

A partir da identificação dos fatores de risco, é possível prevenir o desenvolvimento de cardiopatias e viver uma vida saudável. Entretanto, o acesso a este nível de informação, bem como à renda necessária para se manter devidamente saudável, não são idealmente distribuídos, fazendo com que uma grande parcela da população não tenha acesso ao diagnóstico e ao tratamento devido. Dessa forma, a predição do diagnóstico desse grupo de enfermidades é fundamental para evitar milhões de mortes e as tecnologias de Data Science são uma ferramenta poderosa para atingir esse objetivo.

No atual momento, o aprendizado de máquina é muito utilizado para classificação de dados. Portanto, a partir de dados disponibilizados publicamente no repositório de aprendizagem de máquina UCI [1], construímos um método de classificação utilizando o modelo do Classificador Ingênuo de Bayes (Naïve Bayes) na predição da classe de dados de pacientes que podem sofrer de doenças cardíacas ou não.

II. Objetivos

Como citado anteriormente, dados da OMS mostram que as mortes por doenças relacionadas ao coração crescem cada vez mais a cada ano. Em vista disso, o intuito deste trabalho acadêmico foi construir um sistema capaz de ajudar a reduzir esse excesso de mortes. Portanto, para atingir esse objetivo, criamos de um algoritmo de predição de sinais que podem indicar possíveis cardiopatias em pessoas.

O sistema construído aqui tem o objetivo de identificar características e padrões associados às doenças cardiovasculares em uma base de dados do repositório UCI [1]. Então, dados coletados de pacientes que tiveram doenças cardíacas foram analisados utilizando o Classificador Ingênuo de Bayes a partir de recursos e bibliotecas de aprendizagem de máquina para identificar esses padrões e, a partir deles, foi possível inferir um diagnóstico positivo ou negativo de uma doença cardíaca.

Sendo assim, esse modelo tem o objetivo de ser uma ferramenta capaz de ajudar profissionais da saúde a tomar decisões clínicas mais rápidas e precisas do que os sistemas tradicionais de apoio podem oferecer, a fim de diminuir os números tão abundantes de doenças cardiovasculares no mundo.

III. Justificativa

Nos últimos anos, com o crescimento da produção de alimentos ultra processados e o aumento da jornada de trabalho, surge uma nova tendência de hábitos relacionados à saúde. Dessa forma, cresce a cada ano o número de pessoas que desenvolvem fatores de risco de adquirir doenças cardiovasculares, como pressão alta, diabetes e sedentarismo.

Mesmo assim, o acesso ao diagnóstico e ao tratamento para essas doenças ainda enfrenta diversos desafios. Grupos de classes socioeconômicas distintas têm taxa de mortalidade por doenças cardiovasculares diferentes, o que também se observa quando se considera gênero ou raça diferentes, ressaltando a desigualdade no acesso ao tratamento devido [6]. Diante disso, é de fundamental importância o uso de uma ferramenta que possa identificar os fatores de risco e prever a incidência de doenças cardíacas de modo a direcionar o paciente ao tratamento devido e tornar o acesso a ele menos desigual.

IV. Base de Dados

A fim de desenvolver um sistema de predição, será utilizada a base de dados "Heart Disease Data Set" [1] que contém dados coletados em 1988 de 303 pacientes dos institutos e hospitais: Hungarian Institute of Cardiology, Budapeste; University Hospital, Zurique; University Hospital, Basel, Suíça; V.A. Medical Center, Long Beach e Cleveland Clinic Foundation.

Essa base de dados possui 76 atributos, mas apenas 14 deles são utilizados:

Tabela I
Descrição dos parâmetros da base de dados

Atributo	Descrição
age	Idade em anos
sex	Valor 1: masculino. Valor 0: feminino
cp	Tipo da dor no peito. Valor 1: angina típica. Valor 2: angina atípica. Valor 3: dor não-anginosa. Valor 4: assintomático
trestbps	Pressão sanguínea em repouso medida em mmHg
chol	Colesterol sérico em mg/dl
fbs	Nível de açúcar no sangue em jejum >120mg/dl. Valor 1: verdadeiro. Valor 0: falso
restcg	Resultado de eletrocardiografia em repouso. Valor 0: normal. Valor 1: tem anormalidade ST-T. Valor 2: demonstra hipertrofia ventricular esquerda (LVH)
thalach	Frequência cardíaca máxima
exang	Angina induzida por exercício. Valor 1: sim. Valor 0: não
oldpeak	Depressão do segmento ST induzida por exercício em relação ao repouso
slope	Inclinação do oldpeak.Valor 0: ascendente. Valor 1: plano. Valor 2: descendente
ca	Número de vasos sanguíneos
thal	Determina o quão bem o sangue flui pela musculatura do coração. Valor 3: normal. Valor 6: fixed defect. Valor 7: reversable defect
num	Diagnóstico de doença cardíaca, é o estreitamento das artérias dado pelo resultado de uma angiografia. Valor 0: < 50% diameter narrowing. Valor 1: > 50% diameter narrowing

V. Análise Exploratória dos Dados

Antes de utilizar os dados para fazer inferências, é necessário compreender a forma em que eles estão dispostos. Para isso, uma análise exploratória dos dados foi realizada utilizando ferramentas de bibliotecas da linguagem Python da seguinte forma:

A. *Entendendo as Variáveis*

*1) **Tipos das Variáveis**:* Primeiro, foi preciso entender o que as variáveis representam, nomear os parâmetros da base de dados e, a partir da visualização das informações dos dados, como as colunas possuíam tipos diferentes, todos os dados passaram a ser do tipo "float"para facilitar sua manipulação.

Tabela II
Tratamento dos tipos de dado das variáveis

Variável	Tipo Antes	Tipo Depois
age	float64	float32
sex	float64	float32
cp	float64	float32
trestbps	object	float32
chol	object	float32
fbs	object	float32
restcg	object	float32
thalach	object	float32
exang	object	float32
oldpeak	object	float32
slope	object	float32
ca	object	float32
thal	object	float32
num	int64	float32

*2) **Remoção de outliers**:* Ao visualizar as informações da tabela em diagramas de bloco, percebe-se que várias colunas possuíam dados com valores muito diferentes dos demais que poderiam causar anomalias nos resultados obtidos, são os chamados outliers. Para resolver esse problema, todas as linhas da tabela que estavam a uma distância superior a 3 desvios padrão da média foram removidas.

*3) **Remoção de valores inválidos**:* No entanto, mesmo após essa remoção, algumas colunas ainda possuíam valores inválidos. Por exemplo, a coluna que indica o colesterol total ("chol") possuía diversas linhas com valor "0", no entanto, não é possível ter valores tão baixos de colesterol. Por isso, essas linhas inválidas foram substituídas pela média de todos os valores da coluna.

Além disso, a coluna que indicava o diagnóstico para doenças cardíacas ("num") deveria possuir apenas valores binários (0 ou 1), mas havia dados presentes nela que não correspondiam a essa característica. Então, todas as linhas com esses valores inválidos foram removidas.

*4) **Preenchimento de valores ausentes**:* Ademais, percebe-se que várias colunas possuem valores faltando. Para tratar esse problema, os dados foram separados em dois tipos: categóricos e numéricos.

- **Dados numéricos:** foram substituídos pela média dos dados da coluna.
- **Dados categóricos:** foram substituídos pelo valor mais frequente da coluna (moda).

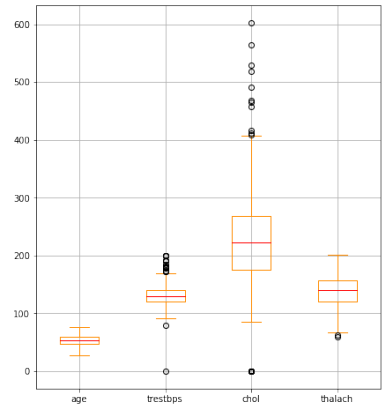

Figura 1. Diagrama de Bloco antes da remoção de outliers

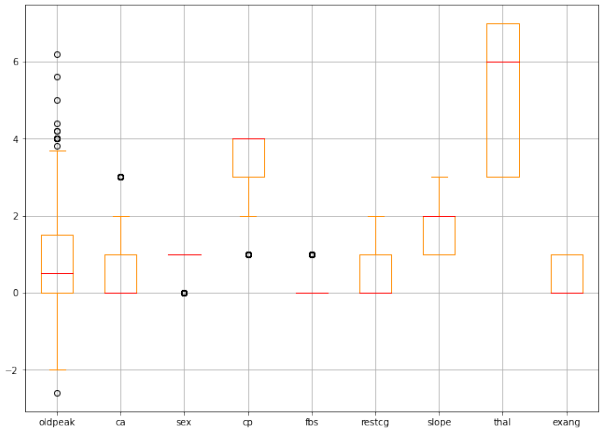

Figura 2. Diagrama de Bloco antes da remoção de outliers

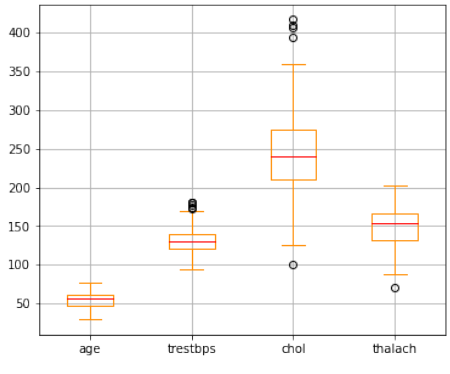

Figura 3. Diagrama de Bloco após a remoção de outliers

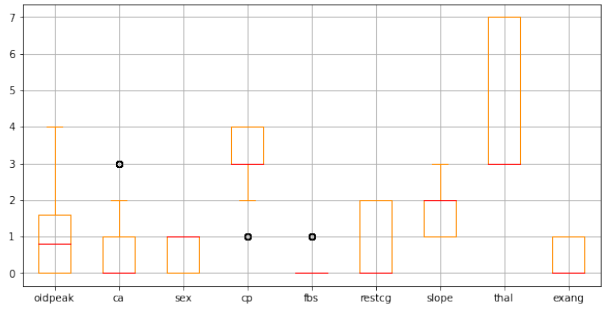

Figura 4. Diagrama de Bloco após a remoção de outliers

B. Representação Gráfica

Os dados numéricos puderam ser representados graficamente no intuito de visualizar suas distribuições. Por exemplo, alguns gráficos possuem uma distribuição gaussiana, como a idade (age), a frequência cardíaca máxima (thalach), entre outros, seja ela simétrica ou não.

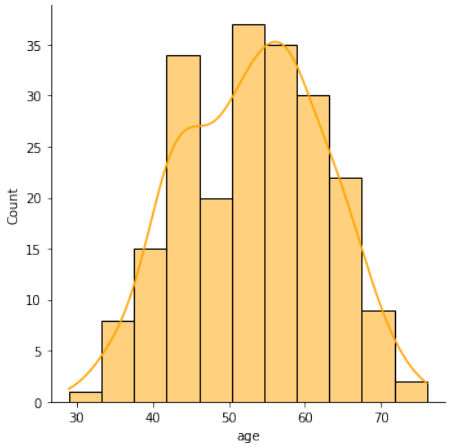

Figura 5. Distribuição de Idade (age)

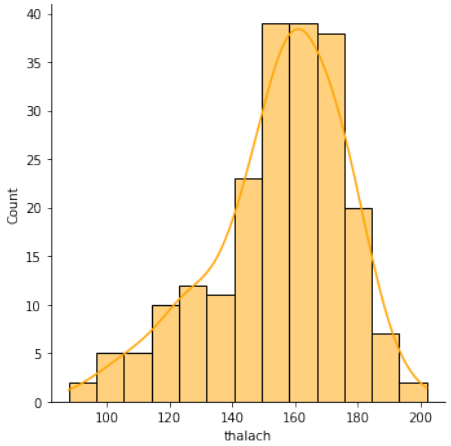

Figura 6. Distribuição de Frequência Cardiáca Máxima (thalach)

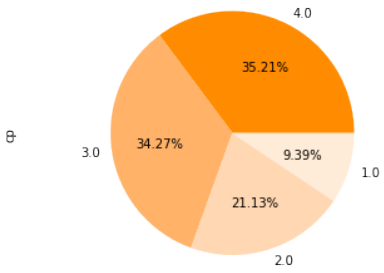

Figura 7. Representação Gráfica por tipo de dor no peito (cp)

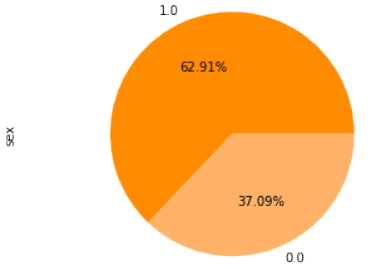

Figura 8. Representação Gráfica por Sexo (sex)

Já os dados categóricos puderam ser representados de maneira mais intuitiva em gráficos do tipo Pie. Por exemplo, a figura 7, que indica o tipo da dor no peito: 1, angina típica; 2, angina atípica; 3, dor não-anginosa; 4, assintomático. Além da figura 8 que indica o sexo da pessoa em questão, 1 para masculino e 0 para feminino.

C. *Análise Estatística*

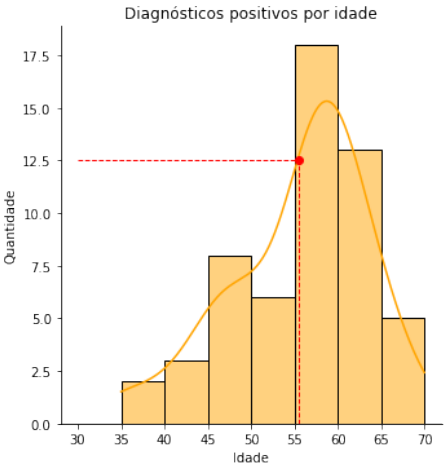

Figura 9. Diagnósticos positivos por idade

Após a representação de alguns gráficos, uma análise estatística foi realizada a partir dos dados para visualizar melhor as características presentes nos indivíduos com diagnóstico positivo para doenças cardíacas. Para isso, foram calculadas as médias de algumas variáveis.

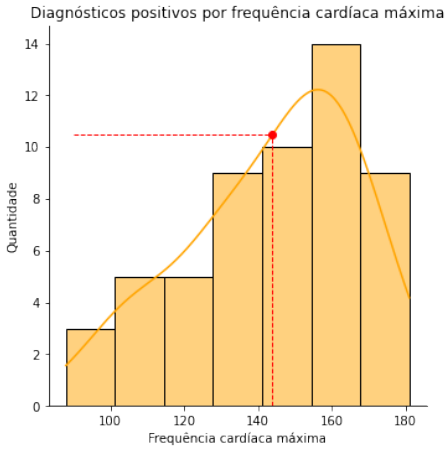

Figura 10. Diagnósticos positivos por Frequência Cardíaca Máxima

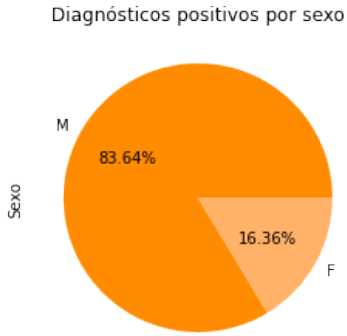

Figura 11. Diagnósticos positivos por sexo

Por exemplo, na figura 9, podemos identificar que a média de idade das pessoas diagnosticadas com doenças cardíacas é de 55.49 anos.

Além disso, na figura 10, nota-se que a média da frequência cardíaca máxima dentre os que têm diagnóstico positivo é de 143.75 bpm.

A figura 10 representa os diagnósticos positivos por sexo, o que nos leva a afirmar que mais de 80% dos pacientes com doenças cardíacas observados são homens.

VI. Implementação do Classificador Ingênuo de Bayes

A. *Classificador Ingênuo de Bayes*

O Classificador Ingênuo de Bayes é um método de classificação baseado no Teorema de Bayes, o principal conceito da probabilidade condicional. É um método robusto e de alto desempenho por sua simplicidade. Como o seu nome sugere, o algoritmo supõe, ingenuamente, que todas as variáveis no conjunto de dados são independentes umas das outras, ou seja, que não há correlação entre elas. O modelo é de fácil implementação e pode ser praticado em grandes data sets, por este motivo é amplamente utilizado. A base desse classificador é o Teorema de Bayes, dado pela Equação (1):

$$P(C|A) = \frac{P(C)P(A|C)}{P(A)} \quad (1)$$

Onde:

- A: atributos;
- C: classe;
- $P(C|A)$: probabilidade condicional de C acontecer dado que A ocorreu;
- $P(A|C)$: probabilidade condicional de A acontecer dado que C ocorreu;
- $P(C)$: probabilidade do evento C;
- $P(A)$: probabilidade do evento A;

Como estamos tratando conjuntos de dados reais, as informações coletadas têm várias singularidades, tornando os cálculos bastante complicados. Portanto, a independência entre as variáveis nesse modelo é utilizada para separar esses vários detalhes e tratá-los cada variável como independente.

A forma de aplicação do classificador Ingênuo de Bayes depende diretamente de como as variáveis em questão estão distribuídas.

Após visualizar em gráficos a distribuição dos dados, percebemos que alguns deles têm Distribuição Normal e, como temos características reais sendo retratadas por variáveis aleatórias contínuas, assumirmos que todos os valores são distribuídos normalmente. Dessa forma, usamos a distribuição Gaussiana para implementar o Naïve Bayes e checar padrões no data set, e assim descobrir a probabilidade de acerto na predição de um diagnóstico.

B. *Distribuição Gaussiana*

As distribuições de probabilidade de grande parte dos fenômenos naturais podem ser muito bem representadas pela Distribuição Gaussiana (ou Normal). Portanto, se temos características reais sendo retratadas por variáveis aleatórias contínuas, assumiremos que esses valores são distribuídos normalmente. Dessa forma, a utilização do Classificador Ingênuo de Bayes em Distribuições Gaussianas é descrita abaixo pela equação (2):

$$p(a_i|c_j) = \frac{1}{\sqrt{2\pi}\sigma_j} e^{-\frac{(a_i-\mu_j)^2}{2\sigma_j^2}} \quad (2)$$

Onde:

- $p(a_i|c_j)$: probabilidade condicional de a_i acontecer dado que c_i ocorreu;
- a_i: i-ésimo atributo;
- c_j: j-ésima classe;
- μ: média;
- σ^2: variância;

VII. Análise de Resultados: conceitos preliminares

A. *Presença de overfitting*

Um dos problemas mais recorrentes no treinamento de modelos é o overfitting. Overfitting ocorre quando o modelo aprende demais com os dados de treino e acaba se tornando adequado apenas para essa base de dados. Em um primeiro momento, precisamos saber se o modelo não sofre de overfitting, para isso, utilizamos a validação cruzada, uma técnica que consiste em dividir a base de dados para treinamento em k bases menores e treinar k modelos com cada uma com uma dessas bases. É uma técnica útil para checar se o modelo sofre de overfitting quando a performance dele é anormalmente alta.

B. *Métricas e indicadores*

A avaliação da performance foi dada através de métricas e indicadores bem estabelecidos para problemas de classificação: o recall e a precisão. Para isso, utilizamos a matriz de confusão, representada na figura 12, uma ferramenta onde as linhas são os rótulos (classes) do seu conjunto de dados de teste e as colunas são as classes que seu modelo previu. Essa ferramenta nos permite ver em detalhes o número de ocorrências onde o modelo confundiu classes e onde acertou.

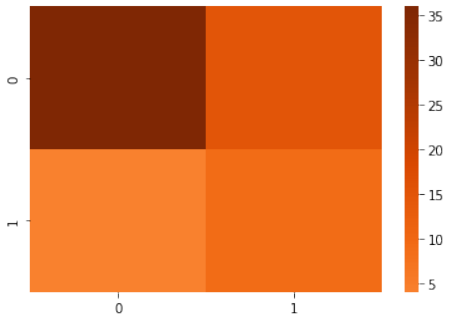

Figura 12. Matriz de Confusão do Classificador

Tabela III
Saída da Matriz de Confusão

	Previsão de Diagnóstico Negativo	**Previsão de Diagnóstico Positivo**
Verdadeiro Negativo	36	15
Verdadeiro Positivo	4	9

Seja um caso e a classe que ele pertence, temos as seguintes saídas da matriz de confusão (Tabela III):

- **Verdadeiro positivo** — Quando se prevê que um caso pertence a sua classe.
- **Verdadeiro negativo** — Quando se prevê que um caso não pertence a uma classe que não é sua classe.
- **Falso negativo** — Quando se prevê que um caso não pertence a uma classe que é sua classe.
- **Falso positivo** — Quando se prevê que um caso pertence a uma classe que não é sua classe.

Com essas saídas, torna-se possível calcular os dois indicadores citados acima:

- **Precisão** — É a razão entre os verdadeiros positivos e a soma entre os verdadeiros positivos e os falso positivos. A precisão mede a porcentagem de acertos nas classes positivas.

- **Recall** — É a razão entre os verdadeiros positivos e a soma entre os verdadeiros positivos e falso negativos. O recall mede a porcentagem de acertos nas classes corretamente previstas.
- F_1score - é uma que combina os dois indicadores através de uma média harmônica dada pela equação (3) abaixo:

$$F_1score = \frac{2}{\frac{1}{\text{Precisão}} + \frac{1}{\text{Recall}}} \tag{3}$$

F_1score só será alto se a precisão e o recall forem altos.

Com esses indicadores e ferramentas, pode-se analisar os resultados e calibrar o modelo.

Tabela IV
INDICADORES

	Precisão	Recall	f1-score	Suporte
Negativo (0)	0.90	0.71	0.79	51
Positivo (1)	0.38	0.69	0.49	13
Acurácia	-	-	0.70	64
Média Macro	0.64	0.70	0.64	64
Média Ponderada	0.79	0.70	0.73	64

VIII. Experimentos e Resultados

A. *Considerações iniciais*

Visto que nosso modelo tem como fim classificar se um indivíduo tem doença do coração, sabemos que prever um diagnóstico negativo quando na verdade se tem um diagnóstico positivo - um falso negativo - é muito mais prejudicial que prever um diagnóstico positivo quando na verdade há um diagnóstico negativo, ou seja, um falso positivo. Dentro desse contexto, temos que nosso objetivo é minimizar os falsos negativos. Portanto, a métrica que será o foco da nossa análise será o recall, que deve ser usado quando os falsos negativos são mais prejudiciais que os falsos positivos.

B. *Primeiro experimento*

No primeiro experimento, o modelo foi treinado e foi avaliada sua performance para ser usada como parâmetro de comparação com os próximos testes. foram obtidos os seguintes resultados:

Tabela V
SAÍDA DA MATRIZ DE CONFUSÃO: PRIMEIRO EXPERIMENTO

	Previsão de Diagnóstico Negativo	Previsão de Diagnóstico Positivo
Verdadeiro Negativo	36	15
Verdadeiro Positivo	4	9

Tabela VI
INDICADORES: PRIMEIRO EXPERIMENTO

	Precisão	Recall	f1-score	Suporte
Negativo (0)	0.90	0.71	0.79	51
Positivo (1)	0.38	0.69	0.49	13
Acurácia	-	-	0.70	64
Média Macro	0.64	0.70	0.64	64
Média Ponderada	0.79	0.70	0.73	64

C. *Segundo Experimento*

No segundo experimento, foi feita a estratificação dos dados pelos rótulos das classes, visando observar o efeito de um conjunto de treinamento e de teste balanceados.

Nesse experimento, podemos ver a influência da estratificação. Temos o aumento do recall em detrimento da precisão no grupo mais hegemônico (diagnóstico negativo) e o aumento da precisão em detrimento do recall na classe de diagnóstico positivo.

Nesse experimento, foi observado uma diminuição de falsos negativos. Após a estratificação, obteve-se os seguintes resultados:

Tabela VII
SAÍDA DA MATRIZ DE CONFUSÃO: SEGUNDO EXPERIMENTO

	Previsão de Diagnóstico Negativo	Previsão de Diagnóstico Positivo
Verdadeiro Negativo	35	12
Verdadeiro Positivo	8	9

Tabela VIII
INDICADORES: SEGUNDO EXPERIMENTO

	Precisão	Recall	f1-score	Suporte
Negativo (0)	0.81	0.74	0.78	47
Positivo (1)	0.43	0.53	0.47	17
Acurácia	-	-	0.69	64
Média Macro	0.62	0.64	0.63	64
Média Ponderada	0.71	0.69	0.70	64

D. *Terceiro Experimento*

No terceiro experimento, foi utilizado o algoritmo Boruta para a escolha das variáveis. Esse algoritmo baseia-se no modelo de florestas aleatórias e nos retorna um conjunto da melhores variáveis para se usar.

Foi limitada a altura da árvore para 3, para evitar overfitting, que é relativamente comum em modelos baseados em árvores.

O algoritmo retornou ”age”,”cp”,”thalach”,”exang”, ”oldpeak”e ”thal”.

Nesse experimento, o modelo obteve a maior quantidade de falsos negativos dos experimentos realizados. O baixo aumento do índice de precisão foi desproporcional a diminuição do

recall na classe de diagnóstico negativo e o aumento do recall na classe de diagnóstico positivo foi irrisório em relação a diminuição da precisão.

Tabela IX
SAÍDA DA MATRIZ DE CONFUSÃO: TERCEIRO EXPERIMENTO

	Previsão de Diagnóstico Negativo	**Previsão de Diagnóstico Positivo**
Verdadeiro Negativo	34	17
Verdadeiro Positivo	6	7

Tabela X
INDICADORES: TERCEIRO EXPERIMENTO

	Precisão	**Recall**	**f1-score**	**Suporte**
Negativo (0)	0.85	0.67	0.75	51
Positivo (1)	0.29	0.54	0.38	13
Acurácia	-	-	0.64	64
Média Macro	0.57	0.60	0.56	64
Média Ponderada	0.74	0.64	0.67	64

E. Quarto experimento

No quarto experimento, os dados do experimento anterior foram estratificados antes do modelo ser treinado.

Nesse experimento, tivemos a menor quantidade de falsos negativos. O recall do diagnóstico negativo aumentou em detrimento da precisão, ocorreu o inverso na classe de diagnóstico positivo, onde a precisão aumentou em detrimento do recall.

Tabela XI
SAÍDA DA MATRIZ DE CONFUSÃO: QUARTO EXPERIMENTO

	Previsão de Diagnóstico Negativo	**Previsão de Diagnóstico Positivo**
Verdadeiro Negativo	40	7
Verdadeiro Positivo	10	7

Tabela XII
INDICADORES: QUARTO EXPERIMENTO

	Precisão	**Recall**	**f1-score**	**Suporte**
Negativo (0)	0.80	0.85	0.82	47
Positivo (1)	0.50	0.41	0.45	17
Acurácia	-	-	0.73	64
Média Macro	0.65	0.67	0.64	64
Média Ponderada	0.72	0.73	0.73	64

IX. CONCLUSÕES E DISCUSSÕES

A partir da análise dos resultados, percebe-se que o modelo não apresenta uma performance boa quando prediz diagnósticos positivos. Contudo, para a previsão de diagnósticos negativos para doenças cardiovasculares, ele tem um desempenho com uma precisão de 80% e um Recall de 85%. Ou seja, é um sistema eficaz em afirmar a ausência de cardiopatias do que em afirmar sua presença.

Portanto, embora o objetivo de construir um Classificador que prediz se uma pessoa tem uma doença cardíaca não foi alcançado, conseguimos um modelo que é capaz de afirmar o contrário, ou seja, indicar que um paciente não possui doença cardiáca a partir dos dados recebidos.

Assim, visto que receber um diagnóstico negativo para doenças cardíacas sendo verdadeiro positivo (falso negativo) é muito mais prejudicial do que receber um diagnóstico positivo sendo verdadeiro negativo (falso positivo) dentro do nosso contexto, tem-se que o modelo, por ter um recall de 85% para diagnóstico negativo, explicita a minimização de falso negativos.

Dessa forma, mesmo que o objetivo inicial não tenha sido atingido, o classificador consegue predizer bem casos que não são cardiopatias, evitando a maior parte dos falso negativos.

REFERÊNCIAS

[1] Andras Janosi, William Steinbrunn, Matthias Pfisterer, Robert Detrano, "Heart Disease Data Set", UCI — Machine Learning Reporitory. [Online]. Disponível: https://archive.ics.uci.edu/ml/datasets/heart+disease

[2] "Cardiovascular disease", Nhs.uk. [Online]. Disponível em: https://www.nhs.uk/conditions/cardiovascular-disease/. [Acessado: 09-nov-2021].

[3] N. Salmi and Z. Rustam, "Naïve Bayes Classifier Models for Predicting the Colon Cancer," vol. 546, p. 52068, Jun. 2019, doi: 10.1088/1757-899x/546/5/052068.

[4] "The top 10 causes of death", Who.int. [Online]. Disponível em: https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death. [Acessado: 09-nov-2021].

[5] A. Pattekari, S.A.; Parveen, "Prediction system for heart disease using Naïve Bayes," Int. J. Adv. Comput. Math. Sci., vol. 3, no. 3, pp. 290–294, 2012.

[6] G. A. Mensah, "Eliminating disparities in cardiovascular health: six strategic imperatives and a framework for action: Six strategic imperatives and a framework for action", Circulation, vol. 111, no 10, p. 1332–1336, 2005.

[7] A. Geron, Hands-on machine learning with scikit-learn, keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems, 2nd ed. Sebastopol, CA: O'Reilly Media, 2019.

[8] P. Banerjee, "Naive Bayes Classifier in Python", Kaggle.com, 28-ago-2020. [Online]. Disponível em: https://www.kaggle.com/prashant111/naive-bayes-classifier-in-python. [Acessado: 09-nov-2021].

[9] Wikipedia contributors, "Bernoulli distribution", Wikipedia — The Free Encyclopedia, 01-nov-2021. [Online]. Disponível em: https://en.wikipedia.org/w/index.php?title=Bernoulli_distribution&oldid=1053087848. [Acessado: 09-nov-2021].

[10] Wikipedia contributors, "Multinomial distribution", Wikipedia — The Free Encyclopedia, 04-nov-2021. [Online]. Disponível em: https://en.wikipedia.org/w/index.php?title=Multinomial_distribution&oldid=1053486558. [Acessado: 09-nov-2021].

[11] "Pandas documentation — pandas 1.3.5 documentation", Pydata.org. [Online]. Disponível em: https://pandas.pydata.org/docs/. [Acessado: 18-dez-2021].

[12] "seaborn: statistical data visualization — seaborn 0.11.2 documentation", Pydata.org. [Online]. Disponível em: https://seaborn.pydata.org/. [Acessado: 18-dez-2021].

[13] "Numpy and Scipy Documentation — Numpy and Scipy documentation", Scipy.org. [Online]. Disponível em: https://docs.scipy.org/doc/. [Acessado: 18-dez-2021].

[14] "Matplotlib documentation — Matplotlib 3.5.1 documentation", Matplotlib.org. [Online]. Disponível em: https://matplotlib.org/stable/index.html. [Acessado: 18-dez-2021].