Classificador Ingenuo de Bayes para predizer o estado cardiovascular de um indivíduo com base em indicadores chaves

1st Rafael do Nascimento Moura
Centro de Informática - Cin
Universidade Federal de Pernambuco
Recife, Brasil
rnm4@cin.ufpe.br

2nd Paulo Sérgio Galdino de Souza
Centro de Informática - Cin
Universidade Federal de Pernambuco
Recife, Brasil
psgs@cin.ufpe.br

***Abstract*—Segundo a Organização Pan-Americana da Saúde (OPAS), as doenças cardiovasculares são uma das principais causas de morte nas Américas, levando a óbito dois milhões de pessoas a cada ano. Esse tipo de doença comumente está relacionada a indicadores chaves com a presença ou não de outras doenças como hipertensão, o uso de tabaco, a prática de exercícios, o consumo de álcool, etc. Nesse ínterim, o presente trabalho busca desenvolver uma ferramenta computacional com base no classificador ingênuo de Bayes que seja capaz de predizer o estado cardiovascular de uma pessoa com base nos seus indicadores de sáude.**

***Index Terms*—classificador bayesiano, doenças cardiovasculares**

I. Objetivo

Utilizar um conjunto de dados retirado do website Kaggle para estruturar a construção de uma ferramenta computacional que possa ser capaz de predizer o estado cardiovascular de um indivíduo através de indicadores chave. Nesse contexto, a partir da coleta, análise e adequação dos dados, espera-se ser possível a aquisição de condições suficientes que permitam a aplicação do Classificador Ingênuo de Bayes, de modo a viabilizar a construção dessa ferramenta. Uma vez construída, o objetivo final é auxiliar a identificação e prevenção de doenças cardiovasculares, levando em conta que quanto mais cedo essas doenças são notadas menores são as chances de complicações que podem levar pessoas a óbito.

II. Justificativa

De acordo com o Centro de Controle e Prevenção de Doenças, condições de saúde severa, o estilo de vida, idade, histórico familiar e dentre outros fatores podem aumentar o risco de um ataque cardiaco. A mesma organização, através de uma pesquisa, também afirma que cerca de metade (47%) dos estados unidenses tem ao menos de 1 a 3 fatores de risco para doenças cardiovasculares, como alta pressão arterial, alto colesterol e fumo. De certa forma, alguns fatores de risco não podem ser controlados, como a idade e histórico familiar, mas pode-se, de fato, atentar-se aos sinais e procurar investir em um melhor estilo de vida. Detectar e previnir os fatores que tem o maior impacto nas doenças cardíacas tem um grau de relevância significativa na área de saúde, já que, por evidência, dados do Ministério da Saúde indicam que entre 2010 e 2019 houve um aumento de cerca de 59% nas internações de pessoas de até 40 anos por ataques cardíacos e de 9% nas mortes. Dado o exposto, é factível que o conhecimento prévio do conjunto de sintomas ou aspectos relacionados as doenças cardiovasculares podem salvar vidas e contribuir, de forma expressiva, na diminuição da relação dos casos de doenças cardiovasculares com a taxa de mortalidade de diversas nações. Com efeito, visto que as informações utilizadas na base de dados são de fevereiro de 2022, disponibilizadas por um orgão de saúde (CDC), com a exposição dos fatores mais comuns na aderência de ataques cardíacos, a relevância do projeto, através de ferramentas computacionais como python e suas API's, se caracteriza em permitir que uma das aplicações de métodos de aprendizagem de máquina, como o classificador ingênuo de Bayes, consiga detectar "padrões" a partir de fatores chaves que podem prever a condição de um paciente e predizer se este, é propenso ou não a uma enfermidade cardíaca e, consequentemente, por meio das agências de saúde conseguir detectar e agilizar o tratamento da alteração do sistema cardiovascular em questão e diminuir significativamente as taxas de casos fatais e persistência das alterações, por serem não identificadas/tratadas de forma prévia.

III. Base de Dados

A base de dados utilizada no presente trabalho trata sobre indicadores chave relacionados à doença cardíaca. Os dados tem como base os Estados Unidos onde 47% da população tem pelo menos um fator de risco para doença do coração dentre pressão sanguínea alta, colesterol alto e o ato de fumar.

Originalmente, o conjunto de dados vem do CDC (Centers for Disease Control and Prevention) e é uma parte importante do Behavioral Risk Factor Surveillance System (BRFSS), que realiza pesquisas telefônicas anuais para coletar dados sobre o estado de saúde dos residentes dos EUA. Como o CDC descreve: "Estabelecido em 1984 com 15 estados, o BRFSS agora coleta dados em todos os 50 estados, bem como no Distrito de Columbia e três territórios dos EUA. O BRFSS

completa mais de 400.000 entrevistas com adultos a cada ano, tornando-se o maior sistema de pesquisa no mundo.". O conjunto de dados mais recente (em 15 de fevereiro de 2022) inclui dados de 2020. Ele consiste em 401.958 linhas e 279 colunas. A grande maioria das colunas são perguntas feitas aos entrevistados sobre seu estado de saúde.

As colunas podem ser descritas da seguinte forma:

TABLE I
DESCRIÇÃO DAS VARIÁVEIS

Variável	Descrição
HeartDisease	Respondentes que já relataram ter doença cardíaca coronária (CHD) ou infarto do miocárdio (MI). (Sim / Não)
BMI	Índice de Massa Corporal
Smoking	Respondentes que fumaram pelo menos 100 cigarros na vida (Sim / Não)
AlcoholDrinking	Homens adultos que bebem mais de 14 drinques por semana e mulheres adultas que bebem mais de 7 drinques por semana (Sim / Não)
Stroke	Respondetes que tiveram um AVC (Sim / Não)
PhysicalHealth	Quantos dias durante os últimos 30 dias a saúde física dos respondentes não foi boa? (0-30 dias)
MentalHealth	Quantos dias durante os últimos 30 dias a saúde mental dos respondentes não foi boa? (0-30 dias)
DiffWalking	Os respondentes têm sérias dificuldades para caminhar ou subir escadas? (Sim / Não)
Sex	Feminino ou Masculino
AgeCategory	13 faixas etárias: '18-24', '25-29','30-34','35-39','40-44','45-49','50-54','55 -59','60-64','65-69','70-74','75-79','80 ou mais'
Race	7 grupos raciais 'Branco', 'Preto', 'Asiático', 'Índio Americano/Nativo do Alasca', 'Outro', 'Hispânico'
Diabetic	Status de diabetes dos respondentes: 'Sim', 'Não", 'Não, diabetes limítrofe', 'Sim (durante a gravidez)'
Atividade Física	Respondentes que relataram fazer atividade física ou exercício durante os últimos 30 dias, além de seu trabalho regular. (Sim / Não)
GenHealth	Os respondentes avaliam sua própria saúde em 5 categorias: 'ruim', 'justo', 'bom', 'muito bom', 'excelente'
SleepTime	Em média, quantas horas de sono os respondentes dormem em um período de 24 horas
Asthma	Respondentes que tiveram asma? (Sim / Não)
KidneyDisease	Respondentes que tiveram doença renal, exceto cálculos renais, infecção da bexiga ou incontinência (Sim / Não)
SkinCancer	Respondentes que tiveram câncer de pele? (Sim / Não)

IV. ANÁLISE EXPLORATÓRIA DOS DADOS

A base de dados incial foi retirada do Centers for Disease Control and Prevention, a agência nacional de saúde pública dos Estados Unidos, e contava com certa de 300 variáveis. Antes de ser postada no Kaggle a mesma passou por uma limpeza prévia e a quantidade de vaiáveis foi diminuida para 20. As análises seguintes serão feitas a partir dessa versão.

A. Tipos das Variáveis

As vinte variáveis restantes foram divididas em variáveis categoricas e variáveis numericas. Por sua vez as variáveis categoricas foram divididas em variáveis categoricas binárias, variáveis categóricas ordinárias e variáveis categoricas nominais. Da mesma forma, as variáveis numericas foram divididas em variáveis numericas continuas e discretas.

TABLE II
TIPOS DAS VARIÁVEIS

Tipos	Variáveis
Variáveis Categóricas Binárias	HeartDisease, Smoking, Stroke, Asthma, AlcoholDrinking, KidneyDisease , SkinCancer, PhysicalActivity, DiffWalking
Variáveis Categóricas Ordinárias	AgeCategory, GenHealth
Variáveis Categóricas Nominais	Race, Diabetic
Variáveis Numéricas Contínuas	BMI
Variáveis Numéricas Discretas	PhysicalHealth, MentalHealth, SleepTime

B. Distribuição das Variáveis Númericas - Box Plot

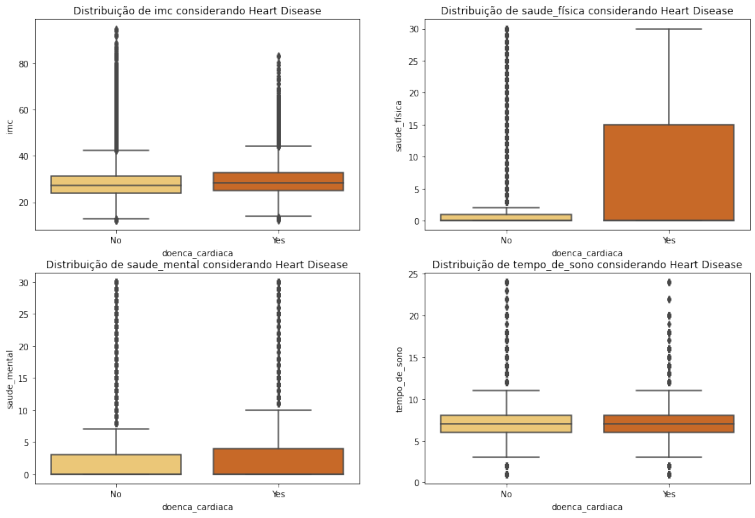

Fig. 1. Box Plot representando a distribuição das variáveis númericas

Uma sequência iterativa é trabalhada para poder representar graficamente todas as variáveis numéricas com um boxplot, considerando que este gráfico permite uma visão mais clara da dispersão dos dados. Aproveita-se para realizar a avaliação de acordo com a variável objetiva 'HeartDisease' de modo

a determinar se existe alguma variável que tenha um perfil discriminante mais marcado que as demais.

Observando os gráficos é possível chegar nas seguintes conclusões:

- A variável BMI apresenta distribuições semelhantes para ambas as categorias na variável alvo. Observa-se um nível médio próximo a 30, com grande número de casos atípicos.
- A variável PhysicalHealth apresenta distribuições diferentes nas duas categorias da variável HeartDisease. No caso da categoria 'Não', observa-se uma baixa dispersão dos resultados (concentrados no nível 0) que são casos que não apresentam dias de saúde física precária. Existe um alto número de valores atípicos que pode ser dado por casos que ainda não tiveram efeitos ou complicações cardíacas. Por outro lado, no caso da categoria 'Sim', observa-se maior dispersão com alta concentração de casos 0 (mediana).
- A variável MentalHealth apresenta distribuições semelhantes para ambas as categorias da variável objetiva com grande número de casos atípicos.
- A variável SleepTime apresenta distribuições muito semelhantes para ambas as categorias.

Analisando agora as variáveis numéricas a partir de histogramas é possível chegar nas mesmas conclusões apresentadas, porém, com algumas adições.

C. Distribuição das Variáveis Númericas - Histogramas

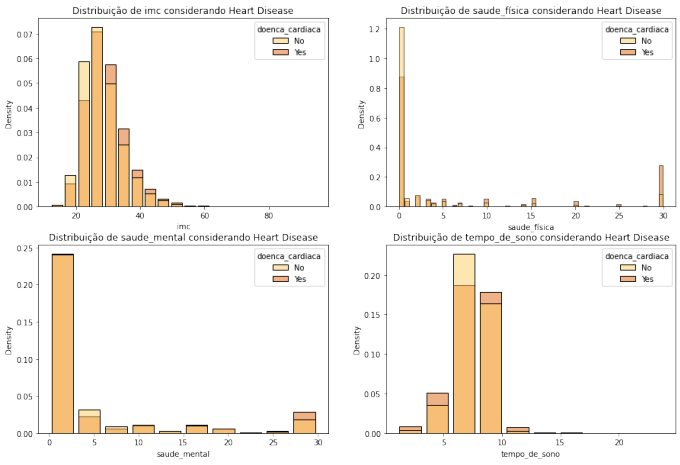

Fig. 2. Histogramas representando a distribuição das variáveis númericas

- Na variável IMC, há uma concentração em valores mais altos para a categoria "Sim" de doença cardíaca
- Nas variáveis saúde física e saúde mental, existe uma alta concentração de valores em 0.
- A variável tempo de sono apresenta distribuições mais concentradas na categoria "Sim" da variável alvo.

D. Distribuição das Variáveis Categóricas - Histogramas

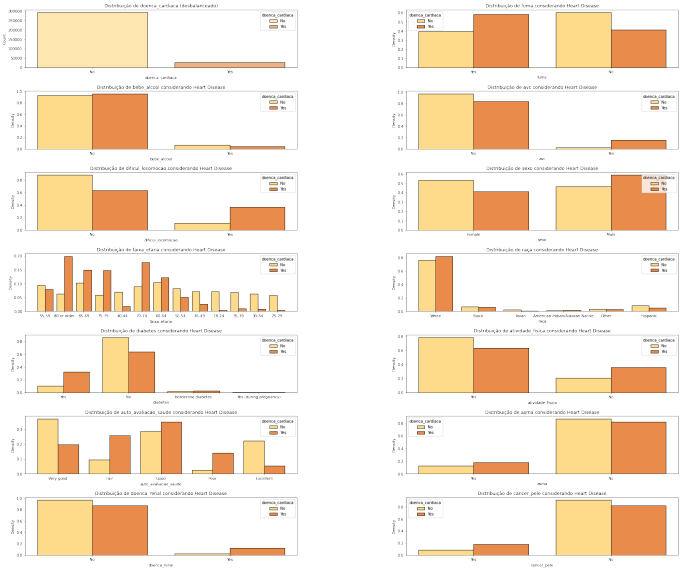

Fig. 3. Histogramas representando a distribuição das variáveis categóricas

Alguns conclusões a partir da análise desses histogramas:

- A maioria dos doentes cardíacos são homens e homens são aproximadamente 1,6 vezes mais propícios a terem problemas cardíacos que mulheres.
- A maioria dos doentes cardíacos fuma, pessoas que fumam são aproximadamente duas vezes mais propícios a terem problemas cardíacos do que pessoas que não fumam.
- A maioria dos doentes cardíacos são brancos, apesar de que há uma amostragem muito maior de pessoas brancas do que de outras raças.
- A maioria das pessoas testadas não apresentam doenças renais, a discrepância é muito grande para afirmar qualquer coisa a partir desses dados, o mesmo.

V. Classificador Ingênuo de Bayes

O classificador multinomial Naïve Bayes é um dos modelos mais populares no aprendizado de máquina. Tomando como premissa a suposição de independência entre as variáveis do problema. O modelo de Naïve Bayes realiza uma classificação probabilística de observações, caracterizando-as em classes pré-definidas. Diante disso, a base de dados utilizada no nosso classificador se trata de dados da pesquisa anual do Centers for Disease Control and Prevention (Uma agência do departamento de saúde dos Estados Unidos) de 2020 de aproximadamente 400 mil adultos entrevistados com perguntas relacionadas ao seu estado de saúde. À vista disso, o principal objetivo ao aplicar a técnica de aprendizagem de máquina Naïve Bayes no projeto em questão, é tornar possível o mapeamento de um paciente aleatório dado uma série de parâmetros chaves que irão prever se o indivíduo em questão é suscetível a ter uma enfermidade cardiovascular ou não. Assim como já explicitado, a base de dados consta com 18 parâmetros

para a realização da associação dos seus respectivos valores com as classes. Após a higienização dos dados e feita a análise exploratória, decidiu-se que, por convenção do algoritmo que foi utilizado, houve a necessidade de transformar todos os tipos de dados das variáveis para apenas um tipo, inteiro. A base de dados consta com os seguintes tipos de variáveis: (9 booleans, 5 strings e 4 decimais). No processo de desenvolvimento do algoritmo, foi decidido que apenas 16 dos 18 atributos categóricos seriam utilizados na aplicação do método. Sendo alguns dos atributos já previamente citados, sendo seus tipos originalmente:

Tipo de variável booleana: Smoking, Stroke, AlcoholDrinking, DiffWalking, Atividade Física, Asthma, Kidney disease e Skincancer;

Tipo de variável String: Sex, Age category, Race, Gen health, Diabetic;

Tipo de variável decimal: PhysicalHealth, Mental Health, SleepTime;

No processo de implementação do algoritmo, foram utilizadas 3 bibliotecas principais: Pandas (para o tratamento da base de dados), Numpy (utilizada para o processamento da separação dos dados e o seu armazenamento em arrays multidimensionais para manipulá-los de melhor forma), Sklearn (para tornar possível as aplicações de aprendizagem de máquina, como a utilização da base de dados no classificador bayesiano, sua acurácia, divisão entre dados de treino e teste, etc), e 3 auxiliares: Sys,Matplotlib e Random (Todas para questões gráficas e tratamento de números e sequências). O método de implementação do algoritmo de Naïve Bayes escolhido foi o método Gaussiano, pois como trabalhamos com muitos tipos dados que são contínuos, foi muito mais viável aplicar um método que assume dados que são descritos por uma distribuição gaussiana sem covariância (independentes) entre as dimensões, em um modelo que pode ser ajustado simplesmente encontrando a média e o desvio padrão dos pontos dentro de cada rótulo, que é tudo o que é necessário para definir tal distribuição, o que permite que as probabilidades das características possam ser assumidas simplesmente como:

$$P(x_i \mid y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}\right)$$

Fig. 4. Calculo das probabilidades dos atributos categóricos.

VI. Experimento

Em questão do experimento, foi realizada uma rápida pesquisa no Centro de Informática da Universidade Federal de Pernambuco com uma amostra de 20 alunos do curso de Engenharia da Computação acerca de suas condições de saúde considerando os 16 parâmetros do nosso modelo do classificador Bayesiano. Por motivos de privacidade, os nomes serão mantidos em anonimato. Os dados relacionados à pesquisa realizada com as 20 pessoas aleatórias: jovens e adultos com a faixa etária de 18-24 anos e 25-29 anos, residentes do Recife e região metropolitana, são retratados por:

Atributos 1								
Entrevistado	Smoking	AlcoholDrinking	Avc	Physhealth	MentalHealth	DiffWalking	Sex	Faixa-etária
#1	Sim	Sim	Não	15	20	Não	Masc.	25-29
#2	Não	Não	Não	0	1	Não	Masc.	18-24
#3	Não	Sim	Não	7	10	Não	Fem.	18-24
#4	Não	Não	Não	5	15	Não	Fem.	18-24
#5	Sim	Sim	Não	0	30	Não	Fem.	18-24
#6	Sim	Sim	Não	2	25	Sim	Masc.	25-29
#7	Não	Não	Não	0	10	Não	Masc.	18-24
#8	Não	Não	Não	14	7	Não	Fem.	18-24
#9	Não	Não	Não	14	14	Não	Masc.	25-29
#10	Sim	Não	Não	10	7	Não	Fem.	18-24
#11	Não	Não	Não	25	14	Não	Masc.	25-29
#12	Não	Não	Não	5	3	Não	Fem.	25-29
#13	Não	Não	Não	2	7	Sim	Masc.	18-24
#14	Sim	Não	Não	1	14	Não	Fem.	18-24
#15	Não	Não	Não	3	5	Não	Fem.	25-29
#16	Não	Sim	Não	0	7	Sim	Fem.	18-24
#17	Não	Não	Não	20	8	Não	Masc.	18-24
#18	Não	Sim	Não	18	10	Sim	Masc.	18-24
#19	Sim	Não	Não	20	14	Sim	Fem.	18-24
#20	Sim	Sim	Sim	15	14	Não	Masc.	18-24

Atributos 2								
Entrevistado	Raça	Diabetes	PhysicalAct.	GenHealth	Sleeptime	Asma	KidneyDisease	SkinCancer
#1	Negro	Sim	4	Normal		Não	Não	Não
#2	Branco	Não	21	Excelente	7	Não	Não	Não
#3	Branco	Não	10	Boa	6	Não	Não	Não
#4	Negro	Não	5	Normal	5	Não	Não	Não
#5	Branco	Não	0	Ruim	5	Sim	Sim	Não
#6	Branco	Sim	0	Ruim	5	Sim	Não	Não
#7	Negro	Sim	12	Normal	10	Sim	Não	Não
#8	Branco	Não	14	Normal	6	Não	Não	Não
#9	Negro	Não	15	Boa	10	Não	Não	Não
#10	Branco	Não	20	Excelente	12	Não	Não	Não
#11	Negro	Não	12	Muito boa	6	Não	Não	Não
#12	Hispânico	Não	4	Normal	6	Não	Não	Não
#13	Branco	Não	4	Ruim	4	Sim	Sim	Não
#14	Branco	Não	6	Normal	6	Não	Não	Não
#15	Branco	Não	10	Muito boa	8	Sim	Não	Não
#16	Negro	Não	8	Muito boa	7	Não	Não	Não
#17	Negro	Não	8	Normal	6	Sim	Não	Não
#18	Hispânico	Não	10	Ruim	4	Sim	Não	Não
#19	Branco	Não	2	Boa	10	Não	Não	Não
#20	Negro	Não	0	Ruim	4	Não	Sim	Não

Fig. 5. Tabela de entrevistados e suas características.

Aproveitando-se dos dados coletados dos alunos entrevistados para aplicar o mapeamento:

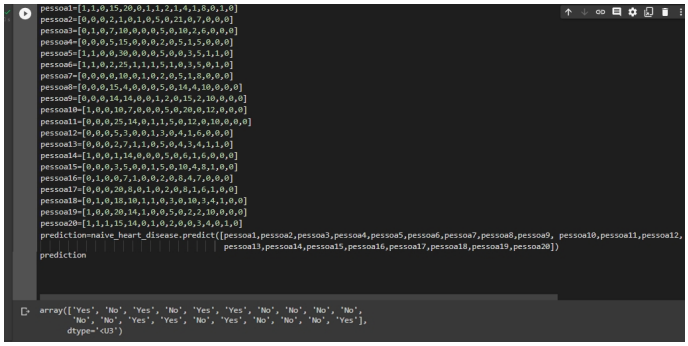

Fig. 6. Mapeamento das classes dos entrevistados dado suas características.

Após a aplicação do método, infere-se que a partir das informações prestadas, os dados foram distribuídos para as classes de tal forma:

Entrevistado	Propenso a enfermidades cardiovasculares (classe)
#1	SIM
#2	NÃO
#3	SIM
#4	NÃO
#5	SIM
#6	SIM
#7	NÃO
#8	NÃO
#9	NÃO
#10	NÃO
#11	NÃO
#12	NÃO
#13	SIM
#14	SIM
#15	NÃO
#16	SIM
#17	NÃO
#18	NÃO
#19	NÃO
#20	SIM

Fig. 7. Mapeamento entre as classes após a aplicação do preditor.

Graficamente,

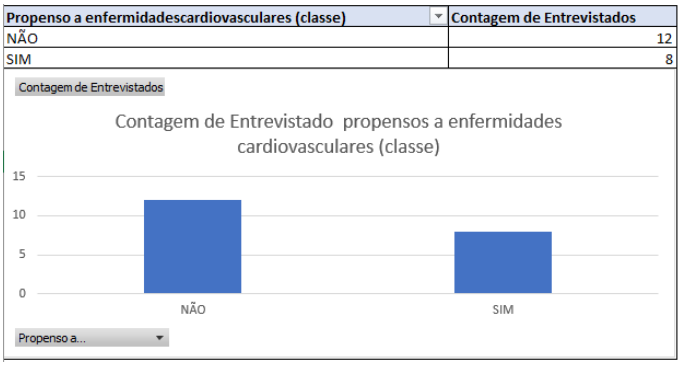

Fig. 8. Resultado das distribuição das classes.

Isso nos mostra que cerca de 60% dos entrevistados não são propensos a doenças cardiovasculares, enquanto 40% podem futuramente apresentar algum quadro de alguma doença cardiovascular. Além disso, a partir da predição realizada, infere-se que a distribuição da classe de pessoas que podem desenvolver essas enfermidades no futuro é bem distribuida entre homens e mulheres (50%) e que pessoas que consomem álcool regularmente/exageradamente, que já fumaram mais de 100 cigarros em toda vida, que dormem poucas horas e possuem asma ou diabetes são muito mais propensas a pertencer a classe "SIM", enquanto os indivíduos que possuem horas regulares de sono, praticam exercícios e tem um bom status de saúde possuem uma probabilidade muito mais alta de pertencer a classe "NÃO".

VII. RESULTADOS

Em virtude da aplicação do método, foi alcançado uma precisão de aproximadamente 85% na predição da probabilidade de quaisquer casos aleatórios:

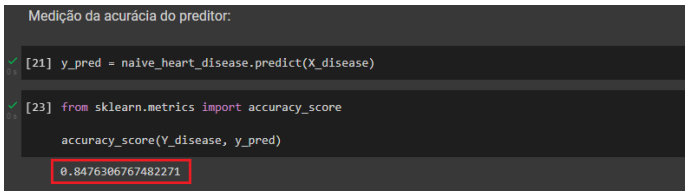

Fig. 9. Trecho de código relativo a acurácia do preditor.

Em nosso modelo, foram utilizados 75% dos dados para treino, enquanto apenas 25% para testes:

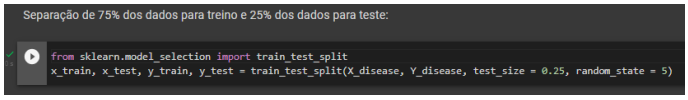

Fig. 10. Divisão dos dados entre treinamento e teste.

Sendo 2 classes possíveis para o mapeamento:

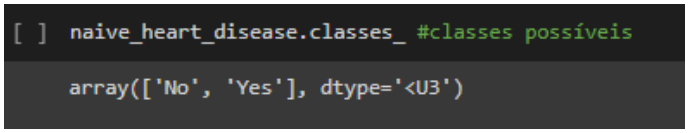

Fig. 11. Classes possíveis a serem mapeadas.

Além disso, pelo fato de a base de dados ser desbalanceada em relação às classes, a quantidade de predições da base de dados foi dada por:

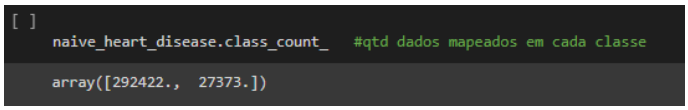

Fig. 12. Distribuição dos casos em cada classe.

O que indica que na classe que prevê "'NÃO", foram mapeados 292.422 casos, do contrário 27.373 "SIM", o que gera uma probabilidade intrínseca a cada classe de:

```
[ ] naive_heart_disease.class_prior_ #prob nao e sim dado a bd

    array([0.91440454, 0.08559546])
```

Fig. 13. Distribuição dos casos em cada classe.

A. *Matriz de confusão*

Foi um método de medição de desempenho utilizado no nosso projeto para a classificação de aprendizado de máquina. A métrica nos auxiliou, de melhor forma, a conhecer o desempenho do modelo de classificação em um conjunto de dados de teste para que os valores verdadeiros e falsos sejam conhecidos. Isso nos tornou claro ao descobrir quantas vezes o nosso modelo deu saída correta ou errada e de que tipo. Por isso, é uma ferramenta muito importante para avaliar modelos de classificação.

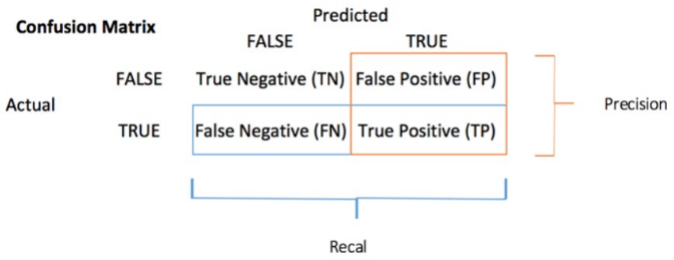

Fig. 14. Subdivisões da matriz de confusão.

Na implementação com a base de dados em questão, a métrica dos dados foi dada por:

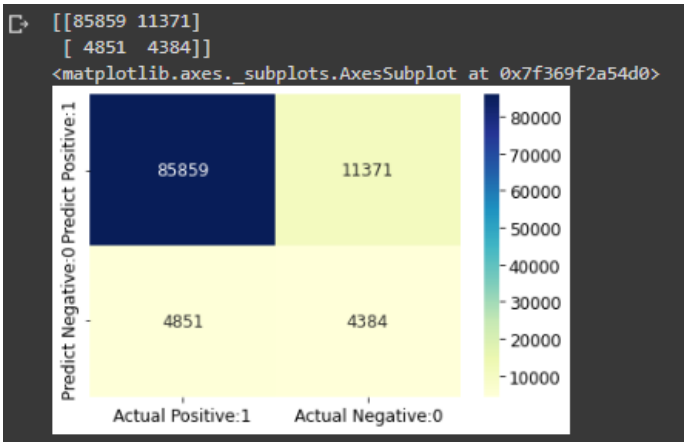

Fig. 15. Matriz de confusão dado a aplicação do classificador bayesiano.

Em que existem 4 tipos líquidos de resultados possíveis:

- TP: True Positive: Valores previstos corretamente previstos como positivos reais;
- FP: False Positive: Os valores previstos previram incorretamente um positivo real. ou seja, valores negativos previstos como positivos;
- FN: False Negative: Valores positivos previstos como negativos;
- TN: True Negative: Valores previstos corretamente previstos como um negativo real;

Na nossa implementação:

- TP: 85859
- FP: 11371
- FN: 4851
- TN: 4384

Acurácia: $(tp + tn)/(tp + tn + fp + fn) = (85859 + 4384)/(85859 + 11371 + 4851 + 4384) = 0.847$

Sendo gerada, as 15 primeiras probabilidades dos casos para as classes (NO,YES) após a aplicação do classificador:

```
y_pred_prob = naive_heart_disease.predict_proba(X_disease)[0:15]

y_pred_prob

array([[1.15366499e-01, 8.84633501e-01],
       [1.21213846e-04, 9.99878786e-01],
       [7.85224179e-02, 9.21477582e-01],
       [8.59441554e-01, 1.40558446e-01],
       [8.85035665e-01, 1.14964335e-01],
       [3.68496132e-01, 6.31503868e-01],
       [5.46484837e-01, 4.53515163e-01],
       [4.22881877e-02, 9.57711812e-01],
       [9.53651379e-05, 9.99904635e-01],
       [8.86712075e-01, 1.13287925e-01],
       [2.09540335e-06, 9.99997905e-01],
       [9.98548165e-01, 1.45183549e-03],
       [4.37366707e-01, 5.62633293e-01],
       [8.23924591e-01, 1.76075409e-01],
       [3.65116032e-01, 6.34883968e-01]])
```

Fig. 16. Probabilidade gerada para os primeiros 15 itens a pertencer as duas classes possíveis. A tabela é totalmente expansível, basta alterar o índice do array para ser capaz de buscar todas as probabilidades para cada caso até o valor limite que engloba todos os casos da base de dados.

Após a aplicação do método, é possível, ainda, realizar as predições. Isto é, dado um paciente aleatório com seus respectivos fatores chaves relacionados à sua saúde, categorizá-lo em uma das classes que indicam se este é propenso a possuir uma doença cardiovascular ou não.

Como já dito previamente, para realizar a predição, todos os tipos de dados foram convertidos para o tipo inteiro, com isso, foi realizada uma associação dos valores presentes na base de dados como segue abaixo:

- Smoking: (Yes=1, No=0);
- Alcohol drinking: (Yes=1, No=0);
- Stroke: (Yes=1, No=0);
- Physical Health: (Foi associado para o próprio valor original, mas em inteiro);
- Mental Health: (Foi associado para o próprio valor original, mas em inteiro);
- Diff Walking: (Yes=1, No=O);
- Sex: (Female=0, Male=1);
- Age Category (18-24 ; 25-29; 30-34, 35-39,40-44,45-49,50-54,55-59,60-64,65-69,70-74,75-79,80 or older) Os valores da faixa de idade são separados em 4 a 4 anos, de (18 anos até 80 or older). Sendo assim , cada valor foi associado como: (18-24 : 0 , 25-29 -¿ 1,30-34 : 2,...,80 or older : 12);

- Race: (White=5,Black=2,American/Indian/alaskan=0,Asian=1, hispanic=3, Other=4);
- Diabetic: (Yes=2,No=0,Borderline diabetes=1 , Diabetes during pregnancy=3);
- Diabetic: (Yes=2,No=0,Borderline diabetes=1 , Diabetes during pregnancy=3);
- PhysicalActivity: (Yes=1, No = 0);
- GenHealth: (Fair=1, Good=2, Very good=4, Poor=3, Excellent=0);
- Sleep Time: (Valor original +1);
- Asthma: (Yes=1,No=0);
- Kidney Disease: (No=0, Yes=1);
- Skin Cancer: (Yes=1, No=0);

É importante frisar que cada parâmetro em questão corresponde a uma posição do array relativo a cada caso que será previsto, cada caso estará dentro de uma lista, isto é, haverá uma lista exclusiva para cada paciente que queira realizar a predição. Para aplicá-la, em termos de código, basta apenas chamar a sua função correspondente:

```
print('Resultado: ' )
prediction=naive_heart_disease.predict([paciente1],[paciente2],[paciente3])
```

Fig. 17. Trecho de código relativo a chamada da função para realizar predições.

E então, dado a sua maior probabilidade, a função retorna o mapeamento para cada um dos casos.

VIII. CONCLUSÃO

O projeto em questão se resumiu a utilizar uma base de dados disponibilizada por uma comunidade on-line de cientistas de dados chamada Kaggle e através de ferramentas computacionais como python e suas bibliotecas, utilizar técnicas de aprendizagem de máquina para higienizar os dados, realizar a análise exploratória e aplicar um algoritmo de classificação probabilística que busca oferecer uma resposta que está entre as opções pré-determinadas em duas classes utilizando as informações somente contidas em uma base de dados. Para isso, foi utilizado o método do classificador probabilístico de Naïve Bayes. A base de dados trata-se de uma pesquisa do ano de 2020 realizada por um órgão de saúde dos Estados Unidos, em que foram entrevistadas mais de 400 mil pessoas acerca de suas condições de saúde. Tendo em vista o conjunto de informações presentes na base de dados, após a higienização, treinamento e conversão dos dados para um só tipo numérico (inteiro), foi possível realizar o mapeamento — buscando a sua maior probabilidade entre pertencer às duas classes possíveis — de um paciente aleatório dado uma série de parâmetros chaves que poderiam prever se o indivíduo em questão é suscetível a ter uma enfermidade cardiovascular ou não, sendo este método realizado através do algoritmo de Naïve Bayes em sua forma Gaussiana, que por utilizar muitas das variáveis em sua forma contínua, considerou que os dados são descritos por uma distribuição gaussiana, isto é, sem covariância (independentes) entre as dimensões, em um modelo que pode ser ajustado simplesmente encontrando a média e o desvio padrão dos pontos dentro de cada rótulo, para assim poder encontrar suas probabilidades características. Após aplicação do método e a escolha de 75% dos dados direcionados para treino e 25% para testes, foi possível alcançar uma acurácia de 84,7% nas predições, dado que existem duas classes muito desbalanceadas entre sí: “YES” - pacientes suscetíveis a uma enfermidade cardiovascular, que na base consta 8,55% do total, aproximadamente 27.373 dos casos, e a classe “NO” - pacientes não suscetíveis a aderir enfermidades cardiovasculares, que nesse caso consta 92.44% do total, aproximadamente 292.422 dos casos. Por fim, através do experimento (uma rápida pesquisa) realizada considerando uma pequena amostra de alunos Centro de Informática, do curso de engenharia da computação da Universidade Federal de Pernambuco, foi possível concluir que a probabilidade entre ambos os sexos feminino e masculino de adquirir uma doença cardiovascular é bem distribuída, além disso, que fatores como o consumo de álcool exagerado ou até mesmo moderado, o fato de ter consumido ao menos 100 cigarros ao longo da vida, sono irregular, diabetes, doenças renais ou asma contribuem para um valor probabilístico consideravelmente maior de possuir problemas cardíacos que aqueles que procuram direcionar o seu estilo de vida a atividades físicas, boas horas de sono e um bom equilíbrio entre saúde mental e física. Visto isso, o método aplicado em questão se torna importante para as mais diversas áreas de saúde, pois é possível, a partir dos indicadores chaves, manter-se em alerta e prognosticar previamente casos ou futuros casos de alterações cardiovasculares, assim sendo identificados de forma prévia, o tratamento seria agilizado, e consequentemente, se buscaria a redução dos grande número de casos que acomete pessoas do mundo todo, pois de acordo com dados da Organização Mundial da Saúde, essas enfermidades matam cerca de 17,7 milhões pessoas todos os anos, sendo somente no Brasil, contabilizados mais 300 mil óbitos anualmente.

REFERENCES

[1] NAIVE Bayes Gaussian. [S. l.], 14 out. 2020. Disponível em: https://scikit-learn.org/stable/modules/generated/sklearn.naive

[2] IMPLEMENTATION of Gaussian Naive Bayes in Python Sklearn. [S. l.], 29 nov. 2021. Disponível em: https://www.analyticsvidhya.com/blog/2021/11/implementation-of-gaussian-naive-bayes-in-python-sklearn/. Acesso em: 21 abr. 2022.

[3] PERSONAL Key Indicators of Heart Disease. [S. l.], 2 jan. 2022. Disponível em: https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease?resource=download. Acesso em: 29 abr. 2022.

[4] GAUSSIAN Naive Bayes. [S. l.], 8 out. 2020. Disponível em: https://iq.opengenus.org/gaussian-naive-bayes/. Acesso em: 10 abr. 2022.

[5] HOW to Develop a Naive Bayes Classifier from Scratch in Python. [S. l.], 7 out. 2019. Disponível em: https://machinelearningmastery.com/classification-as-conditional-probability-and-the-naive-bayes-algorithm/. Acesso em: 8 mar. 2022.

[6] A GENTLE Introduction to Bayes Theorem for Machine Learning. [S. l.], 4 out. 2019. Disponível em: https://machinelearningmastery.com/bayes-theorem-for-machine-learning/. Acesso em: 8 mar. 2022.

[7] ANÁLISE Exploratória. In: CAMPOS, Marcilia Andrade; RÊGO, Leandro Chaves; MENDONÇA, André Feitoza de. Métodos Probabilísticos e Estatísticos com Aplicações em Engenharias e Ciências Exatas. [S. l.: s. n.], 2012. E-book.