

UNIVERSIDADE DE SÃO PAULO
INSTITUTO DE CIÊNCIAS MATEMÁTICAS E DE
COMPUTAÇÃO
DEPARTAMENTO DE CIÊNCIAS DE COMPUTAÇÃO

SCC0276 - APRENDIZADO DE MÁQUINA

PROJETO - PARTE 1

Prof. Fernando Pereira dos Santos
Otávio Cury Pontes - 10716525
Daniel Bernardes Pozzan - 10716608

São Carlos
2022

Conteúdo

1	Introdução	1
2	Descrição e especificações	1
2.1	Dificuldades	3
2.2	Disponibilidade e limpeza de Dados	3
2.3	Métricas para avaliação e comparação dos modelos preditivos .	5
3	Implementação	6
3.1	Análise exploratória	7
3.2	Tratamento de dados	7
3.3	K-nearest neighbours	8
3.4	Naive-Bayes	8
3.5	Comparação	9

1 Introdução

Doenças cardiovasculares (DCV) representam a principal causa de mortes globalmente. Segundo dados da Organização Mundial da Saúde (OMS) [2], DCVs causam 32% de todas as mortes, sendo 85% desses óbitos devido a ataques cardíacos e acidentes vasculares encefálicos. A questão central é que muitas dessas doenças estão associadas a alguns fatores de risco, como tabagismo, dieta alimentar, obesidade, sedentarismo e alcoolismo. Além disso, é imprescindível que DCVs sejam detectadas o quanto antes para que, desse modo, o tratamento necessário seja devidamente aplicado.

A fim de prever a condição de saúde de pacientes, é possível aplicar um modelo de aprendizado de máquina sobre um conjunto de dados e encontrar padrões que relacionam seus hábitos e condições de saúde com a possibilidade do indivíduo portar uma DCV. Para tal, este projeto utilizará um conjunto de dados fornecidos pelo Centro de Controle e Prevenção de Doenças (CDC) dos Estados Unidos. Os dados são coletados anualmente pelo Sistema de Vigilância de Fatores de Risco Comportamental (BRFSS), através de uma entrevista que reúne informações acerca de saúde de residentes dos EUA.

2 Descrição e especificações

O conjunto de dados utilizados neste projeto é composto dos hábitos e informações de saúde, de mais de 400 mil pessoas. Originalmente, nos dados fornecidos pelo BRFSS, havia aproximadamente 300 variáveis a serem consideradas para cada indivíduo. No entanto, para a análise deste projeto, considerou-se apenas as características que estariam mais associadas a doenças cardiovasculares. As tabelas abaixo descrevem as características consideradas para cada paciente.

Dados Qualitativos		
Nome	Descrição	Respostas possíveis
HeartDisease	Entrevistados que já relataram ter doença arterial coronariana (DAC) ou infarto do miocárdio	Verdadeiro ou falso
Smoking	Entrevistados que já fumaram pelo menos 100 cigarros durante toda a vida (Obs.: 5 maços = 100 cigarros)	Verdadeiro ou falso
AlcoholDrinking	Consumo excessivo (Homens adultos que consomem mais de 14 drinks por semana e mulheres adultas que consomem mais de 7 drinks por semana)	Verdadeiro ou falso
Stroke	Entrevistados que já tiveram um acidente vascular encefálico	Verdadeiro ou falso
DiffWalking	Entrevistados que relatam ter dificuldades em caminhar ou subir escadas	Verdadeiro ou falso
Sex	Sexo do entrevistado	Masculino ou feminino
Age	Composto por 14 faixas etárias	65-69, 60-64, entre outros
Race	Composto pela etnia dos entrevistados	Branco, latino, entre outros
Diabetic	Entrevistados que já relataram ter diabetes	Verdadeiro ou falso
PhysicalActivity	Adultos que relataram praticar atividade física ou exercício durante os últimos 30 dias além do trabalho habitual	Verdadeiro ou falso
GenHealth	Estado da saúde geral dos entrevistados	muito boa, boa, entre outros
Asthma	Entrevistados que têm asma	Verdadeiro ou falso
KidneyDisease	Entrevistados que já relataram ter doenças renais (desconsiderando cálculo renal, infecção urinária ou incontinência)	Verdadeiro ou falso
SkinCancer	Entrevistados que já relataram que já tiveram câncer de pele	Verdadeiro ou falso

Tabela 1: Conjunto de dados qualitativos

Dados Quantitativos		
Nome	Descrição	Respostas possíveis
BMI	Índice de massa corporal (IMC)	12 a 94.8
PhysicalHealth	Considerando doenças e lesões físicas, por quantos dias, durante os últimos 30 dias, os entrevistados não estavam saudáveis	0 a 30 dias
MentalHealth	Considerando a saúde mental, por quantos dias, durante os últimos 30 dias, os entrevistados não estavam saudáveis	0 a 30 dias
SleepTime	Em média, quantas horas os entrevistados dormem em um período de 24 horas	1 a 24

Tabela 2: Conjunto de dados quantitativos

2.1 Dificuldades

A dificuldade do problema em questão se dá, justamente, pelo fato de administrar os riscos existentes à vida de um paciente. Apesar de não funcionar como um diagnóstico direto, o método deve relacionar certas características individuais com a possibilidade de um paciente portar alguma doença cardiovascular. De modo geral, o método escolhido deve garantir ao máximo que irá alertar todos os pacientes com risco de portarem DCVs. Qualquer paciente que deixar de ser alertado tem o risco de descobrir a doença sem que haja tempo hábil para tratamento o que, conseqüentemente, pode levá-lo a óbito. Por isso, pode-se definir que o problema exposto consiste em um problema de **classificação**, dado que o modelo irá predizer se um paciente pode possuir ou não, com base nos quesitos expostos, uma doença cardíaca.

2.2 Disponibilidade e limpeza de Dados

O conjunto de dados, disponível na plataforma *Kaggle* [3] em formato *CSV*, é composto por 401.958 linhas (pacientes) e 18 colunas que representam as características já abordadas (9 *booleans*, 5 *strings* e 4 decimais). Tais

dados já estão parcialmente limpos, uma vez que não há dados redundantes ou faltantes dentro do conjunto. No entanto, os dados ainda devem ser tratados com relação a normalização/padronização, pontos aberrantes, desbalanceamento e codificação [4] cujas descrições estão expostas abaixo.

- **Normalização e padronização:** Há dados que variam em intervalos diferentes, como *BMI* e *SleepTime*, *PhysicalHealth* e *MentalHealth*. Tais dados podem ser normalizados a fim de que não haja uma diferença de magnitude, e consequentemente, um prejuízo na análise de tais dados.
- **Pontos aberrantes:** Alguns dos dados da variável *SleepTime*, podem ser considerados aberrantes, como, por exemplo, indivíduos que dormem 1 hora por dia, já que, por diferirem consideravelmente do comum, podem afetar negativamente o modelo de aprendizagem a ser escolhido. Para esta e as demais variáveis quantitativas, é possível definir um intervalo válido para cada uma delas, o que eliminaria tal problema.

SleepTime

On average, how many hours of sleep do you get in a 24-hour period?

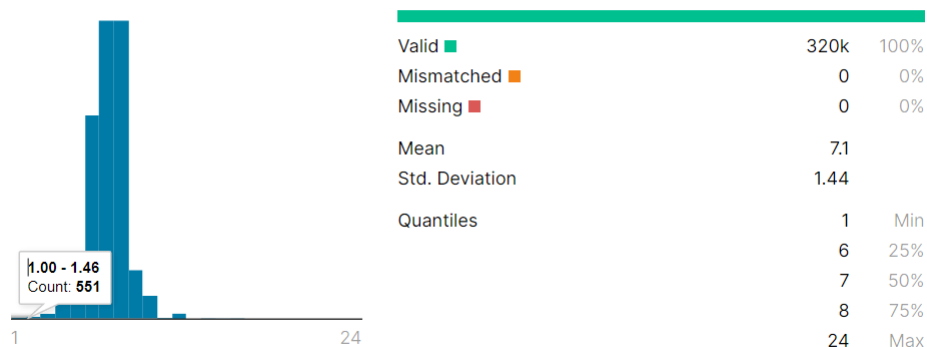


Figura 2.1: Distribuição da variável *SleepTime*.

- **Desbalanceamento:** O desbalanceamento pode ser observado pelo fato de existirem consideravelmente menos indivíduos portadores de

alguma doença cardiovascular. Isso pode ser resolvido escolhendo aleatoriamente uma subamostra de indivíduos saudáveis de mesma magnitude dos pacientes portadores de DCVs.

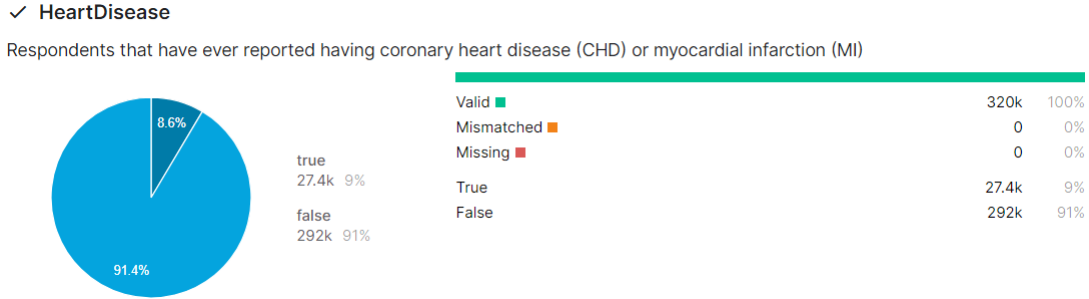


Figura 2.2: Distribuição da variável *HeartDisease*.

- **Codificação:** Algumas variáveis qualitativas ainda podem ser categorizadas numericamente, a fim de facilitar a análise, como é o caso das variáveis *Age* e *GenHealth*.

2.3 Métricas para avaliação e comparação dos modelos preditivos

Como citado anteriormente, é imprescindível que o modelo escolhido apresente o mínimo possível de resultados falso negativos, ou seja, que pacientes portadores de doenças cardiovasculares não sejam alertados. Desse modo, a fim de garantir tal exigência, é possível aplicar algumas métricas sobre os modelos a analisar seus resultados. As métricas [1] que podem ser utilizadas, estão explicitadas abaixo.

	Valor Verdadeiro	
Valor Predito	$Y = 0$	$Y = 1$
$Y = 0$	VN (Verdadeiro Positivo)	FN (Falso Negativo)
$Y = 1$	FP (Falso Positivo)	VP (Verdadeiro Positivo)

Tabela 3: Matriz de Confusão em uma classificação binária.

- **Taxa de falso negativo (TFN):** Para que o modelo seja adequado, a taxa de falso negativo deve ser a menor possível. Ela é obtida por:

$$TFN = \frac{FN}{VP + FN} = 1 - TVP$$

- **Acurácia:** O teste de acurácia, que representa o percentual de quantos elementos foram corretamente identificados, também pode ser aplicado, uma vez que, uma boa precisão geral do modelo também é desejável. A acurácia é obtida através da operação abaixo.

$$Acc = \frac{VP + VN}{VP + VN + FP + FN}$$

- **F1-Score:** Esta métrica relaciona a precisão (valor preditivo positivo), que está focada na taxa de erros por falso positivo, e a revocação (taxa de verdadeiro positivos), que está focada na taxa de erros por falso negativo. É representada por:

$$VPP = \frac{VP}{VP + FP}$$

$$TVP = \frac{VP}{VP + FN}$$

$$F1 = \frac{2 * VPP * TVP}{VPP + TVP} = \frac{2}{\frac{1}{VPP} + \frac{1}{TVP}}$$

3 Implementação

Dadas as deliberações expostas ao decorrer deste relatório, foram implementados alguns modelos de aprendizado de máquina para classificar os dados do conjunto, de forma a prever se uma pessoa pode ou não ter uma doença cardíaca com base nos quesitos expostos. Para a resolução, foi utilizada a ferramenta *Google Colaboratory*, a mesma utilizada nos exercícios e aulas da disciplina.

3.1 Análise exploratória

Antes da aplicação dos modelos, foi feita uma análise exploratória dos dados, a fim de verificar atributos como média, desvio-padrão, valores de mínimo e máximo, além de quartis para as variáveis quantitativas, em conjunto com histogramas e gráficos de frequência, no caso das variáveis qualitativas. Os resultados estão presentes no Notebook entregue junto ao relatório.

3.2 Tratamento de dados

Para a aplicação dos modelos, é necessário um tratamento dos dados de entrada, a fim de obter um modelo mais preciso e diminuir a influência de dados aberrantes e de variáveis contínuas de grande variância.

Em primeiro lugar, as variáveis binárias ('Yes' e 'No') foram transformadas em variáveis booleanas através da ferramenta *LabelEncoder* da biblioteca *sklearn*. Em seguida, foi aplicado *one-hot encoding* para as variáveis qualitativas, de forma a também termos colunas booleanas para estas variáveis.

Em seguida, foi feito um tratamento de *outliers* para as variáveis contínuas, de forma a eliminar a influência estatística que estes dados possam ter no modelo.

Por razão do desbalanceamento da base de dados utilizada, que possui muito mais exemplos de pacientes sem doenças cardiovasculares em relação aos que as têm, é necessário balancear o conjunto de forma a igualar estas quantidades. Para isso, selecionou-se uma parte da base, que, por possuir magnitude de 400 mil linhas, deve ser diminuída para aplicação de modelos de balanceamento de forma a não afetar negativamente o desempenho. Em seguida, foi aplicada a ferramenta *ClusterCentroids* para o balanceamento.

Munidos dos dados balanceados, deve-se fazer a normalização das variáveis contínuas, de forma que todas as variáveis do modelo estão contidas no intervalo $[0, 1]$. Para tal, utilizou-se da ferramenta *MinMaxScaler*, que normaliza os dados com base no modelo de balanceamento mínimo-máximo.

3.3 K-nearest neighbours

Para uma primeira análise, foi aplicado o modelo KNN (*k-nearest neighbours*), utilizando-se da biblioteca *sklearn*, tal qual visto em aula. Para análise dos resultados, o modelo foi executado com $k=1,2,3$ por três vezes cada, cada um utilizando um *random state* diferente aplicado à rotina de *split* dos dados. As tabelas abaixo mostram os resultados obtidos na execução para os valores mencionados de k em relação à acurácia, falsos e verdadeiros positivos/negativos e *F1-score*.

k	1					
Quesito	Acc	FN	FP	TN	TP	F1
\bar{x}	91,87%	144	130	1566	1530	91,78%
σ	0,27%	1	8	11	17,35	0,32%

k	2					
Quesito	Acc	FN	FP	TN	TP	F1
\bar{x}	89,86%	257,67	84	1612	1416,33	89,23%
σ	0,37%	14,57	4,58	13,75	24,01	0,51%

k	3					
Quesito	Acc	FN	FP	TN	TP	F1
\bar{x}	92,99%	109,67	126,67	1569,33	1564,33	92,98%
σ	0,18%	13,05	7,23	15,50	18,90	0,23%

3.4 Naive-Bayes

Com o objetivo de obter uma análise complementar e diminuir o número de falsos negativos, utilizou-se o classificador Naive Bayes, que ignora possíveis correlações entre as variáveis durante o aprendizado. O método foi executado três vezes, em que, para cada execução, foram alterados os valores de *random state* durante o *split* da amostra de dados. A tabela abaixo exhibe as métricas obtidas para o modelo em questão.

Quesito	Acc	FN	FP	TN	TP	F1
Média	93,18%	57,67	172	1498,67	1641,67	93,46%
Desv pad	0,51%	12,86	14,53	19,14	17,79	0,49%

3.5 Comparação

Analisando os resultados obtidos por ambos os métodos, verifica-se que, entre os valores analisados de k no método KNN, o que obteve a melhor acurácia e *F1-score*, como esperado, foi $k = 3$, por analisar mais pontos ao seu redor e, desta forma, determinando melhor a qual categoria o rótulo analisado irá pertencer. A taxa de falsos negativos e positivos se manteve semelhante entre os três, sendo que a menor variação média destes quesitos se encontrou em $k = 1$, o que também é esperado dado que as análises para este caso são mais "binárias" já que se usa somente um vizinho mais próximo para categorizar.

Comparando estes valores com o obtido via Naive-Bayes, verifica-se que este produz bem menos falsos negativos, sendo quase metade do valor encontrado para a melhor versão do KNN, além de possuir uma melhor acurácia e *F1-score* que o KNN também em todos os casos.

Referências

- [1] *Avaliação e Comparação de Modelos*. https://colab.research.google.com/drive/1i8Sr5pH4dmBFVu9Iz8gR9bSn5T6u_9Rp. Acessado em 25/04/2022.
- [2] *Cardiovascular diseases (CVDs)* . [https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-\(cvds\)](https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)). Acessado em 25/04/2022.
- [3] *Personal Key Indicators of Heart Disease*. <https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease>. Acessado em 24/04/2022.
- [4] *Tratamento de Dados*. <https://colab.research.google.com/drive/1xBjemjhHehwfRVfigq3000pNUWXmFsEH>. Acessado em 25/04/2022.