Caracterização de Perfis Clínicos: Aplicação de Algoritmos de Aprendizado de Máquina na Identificação de Indivíduos Saudáveis, Hipertensos e Hipertensos com Doenças Cardiovasculares

Gustavo Costa
G. Costa*
gustavocosta.ds09@gmail.com
Pontifícia Universidade Católica de Minas Gerais
Belo Horizonte, Minas Gerais

ABSTRACT

O campo de Dados, especialmente a Ciência de Dados, é uma área que está em constante expansão. O objetivo primordial dessa ciência é, por meio do estudo dos dados, extrair e formalizar conhecimentos científicos. Logo, é realizado todo o processo de descoberta de conhecimento por meio dos dados para a aplicação deles em modelos de Aprendizado de Máquina para classificar indivíduos como Hipertensos, Hipertensos com Doenças Cardiovasculares ou Saudáveis com, no mínimo, 30 anos de idade. Ademais, a base de dados utilizada no atual trabalho é a Pesquisa Nacional de Saúde do ano de 2019, mais conhecida também como PNS. Essa base de dados consiste em um questionário aplicado pelo Ministério da Saúde do Brasil para uma amostra da população, cerca de 293 mil participantes.

KEYWORDS

mineração de dados, aprendizado de máquina, hipertensão, doenças cardiovasculares, pré-processamento de dados

ACM Reference Format:
Gustavo Costa and G. Costa. 2018. Caracterização de Perfis Clínicos: Aplicação de Algoritmos de Aprendizado de Máquina na Identificação de Indivíduos Saudáveis, Hipertensos e Hipertensos com Doenças Cardiovasculares. In *Woodstock '18: ACM Symposium on Neural Gaze Detection, June 03–05, 2018, Woodstock, NY*. ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/1122445.1122456

Woodstock '18, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-9999-9/18/06…$15.00
https://doi.org/10.1145/1122445.1122456

1 INTRODUÇÃO

A Ciência de Dados, nas últimas décadas, vêm ganhando uma grande influência em todos os aspectos da vida humana no contexto de globalização e de avanços tecnológicos. Essa ciência tem por finalidade analisar múltiplos eventos de diferentes naturezas e as suas ocorrências no cotidiano da vida humana, portanto, ela parte da identificação de variáveis ou fatores que compõe as principais causas do domínio do problema em estudo.

Assim, com a utilização da PNS 2019 no atual trabalho como fonte de dados, o uso da Ciência de Dados e suas técnicas são imprescindíveis para extrair conhecimento sobre a caracterização de perfis clínicos sobre a população brasileira. Essa base de dados contém informações variadas sobre diferentes contextos da vida de uma pessoa: presença de comorbidades, estilos de vida, aspectos socioeconômicos, acesso à serviços básicos, entre muitos outros.

O objetivo deste estudo é aplicar algoritmos de aprendizado de máquina (AM) na identificação de perfis clínicos de pessoas que possuem hipertensão, hipertensão e doenças cardiovasculares ou são saudáveis, isto é, sem nenhuma comorbidade. As doenças cardiovasculares (DCV) são patologias que afetam o sistema circulatório e o sistema cardiovascular como um todo, causando inúmeros problemas na qualidade de vida de um indivíduo e são consideradas um obstáculo árduo para o melhoramento da saúde pública em todo o mundo, sendo a principal causa de morte no Brasil nas últimas 6 décadas.[?] Entre os fatores de risco associados a essas doenças, a Hipertensão Arterial (HA) que constitui outra comorbidade é bastante frequente em países emergentes e subdesenvolvidos, onde a desigualdade social é bem presente na sociedade, como é o caso do Brasil que cerca de 30% da população adulta é atingida.[?] Somado à isso, existem outros fatores que colaboram com a predisposição do surgimento dessas comorbidades, por exemplo, o tabagismo, a falta de atividades físicas, obesidade e a má alimentação.

Assim sendo, o presente artigo visa aplicar algoritmos de AM para caracterizar os perfis clínicos das pessoas que são consideradas como saudáveis, hipertensas ou hipertensas com doenças cardiovasculares simultâneamente. Dessa forma, com o uso desses algoritmos nos dados de saúde pública, busca-se identificar padrões, correlações, tendências que podem passar despercebidas em métodos tradicionais de estudos, por consequência espera-se que o AM aplicado ao contexto da saúde pública seja capaz de gerar novas análises preditivas, previsões de incidências dessas doenças em populações nichadas e com isso melhor direcionar políticas públicas que auxiliem o melhoramento da saúde dos indivíduos como um todo.

2 MATERIAIS E MÉTODOS

2.1 Descrição da Base de Dados

A base de dados utilizada foi a PNS 2019, nela há informações sobre hábitos alimentares, estilos de vida, tabagismo, alcoolismo, aspectos socioeconômicos, acesso à serviços de saúde, saneamento básico, entre outros. Na base disponibilizada pelo Ministério da Saúde, inicialmente, há 293.726 registros e 1.087 atributos. Em seguida, foi selecionada um subgrupo de registros e atributos da base originial, a coleta foi fundamentada com o uso do método CAPTO, que consiste em um processo de descobrimento de conhecimento baseado em conhecimentos tácitos e explícitos sobre o domínio do problema, nesse trabalho o escopo delimitado são fatores associados à HA e a DCV. Portanto, foram selecionados 27 atributos que melhor representam o problema.

Somado à isso, foram utilizadas as instâncias que representam apenas as pessoas com, no mínimo, 30 anos de idade. E ainda, a PNS 2019 foi particionada em três subconjuntos:

- Pessoas Saudáveis: indivíduos que responderam diferente de sim para o diagnóstico das comorbidades HA e DCV e que foram entrevistadas sobre essas perguntas.
- Pessoas Hipertensas: indivíduos que responderam sim para o diagnóstico de HA e responderam não para o diagnóstico de DCV.
- Pessoas Hipertensas com Doenças Cardiovasculares: indivíduos que responderam sim para o diagnóstico de HA e responderam sim para o diagnóstico de DCV.

Para os três subconjuntos, a quantidade de registros são, respectivamente, 112.551, 20.123 e 2.984. Após isso, foi feita a junção dos três subconjuntos em uma base de dados unificada que é a utilizada no pré-processamento de dados e na implementação dos algoritmos de AM, ao todo ela possui 135.658 registros.

De acordo com a figura 1, há 27 atributos selecionados inicialmente e eles configuram-se da seguinte forma:

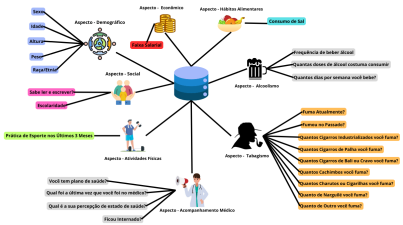

Figure 1: Base de Dados e seus atributos

2.2 Descrição dos Atributos

- **Sexo**: Este atributo indica o gênero da pessoa entrevistada, sendo codificado como um valor categórico. Os valores possíveis são:
  - 1: Homem
  - 2: Mulher
- **Idade**: A idade da pessoa é um valor numérico nominal, representando a quantidade de anos de vida. A idade mínima registrada é de 30 anos e a máxima é de 104 anos.
- **Altura**: Refere-se à altura do indivíduo em centímetros. É um valor numérico nominal, com registros variando de 128 cm até 200 cm.
- **Peso**: Indica o peso corporal do indivíduo em quilogramas, também um valor numérico nominal. O menor valor registrado é 25 kg, enquanto o maior é 170 kg.
- **Raça/Etnia**: Esse atributo representa a raça ou etnia com a qual o entrevistado se identifica, categorizado da seguinte forma:
  - 1: Branco
  - 2: Preto
  - 3: Amarelo
  - 4: Pardo
  - 5: Indígena
  - 9: Ignorado
- **Consumo de Sal**: Esse atributo avalia o nível de consumo de sal do indivíduo, em uma escala ordinal:
  - 1: Muito Alto
  - 2: Alto
  - 3: Adequado
  - 4: Baixo
  - 5: Muito Baixo
- **Praticou Atividade Física nos últimos 3 meses?**: Informa se o indivíduo praticou alguma atividade física nos últimos três meses, com respostas categóricas:
  - 1: Sim
  - 2: Não

- **Fuma atualmente?**: Este atributo pergunta se a pessoa fuma atualmente, indicando o grau de hábito de tabagismo com as seguintes categorias ordinais:
  - 1: Sim, diariamente
  - 2: Sim, menos que diariamente
  - 3: Não fumo atualmente
- **Fumou no passado?**: Pergunta se o indivíduo já fumou anteriormente, sendo categorizado em:
  - 1: Sim, diariamente
  - 2: Sim, menos que diariamente
  - 3: Não
- **Frequência de uso de produtos de tabaco**: Este atributo coleta dados sobre a frequência de uso de produtos específicos de tabaco (cigarros, narguilé, entre outros), com as seguintes opções:
  - 1: Um ou mais por dia
  - 2: Um ou mais por semana
  - 3: Menos que uma vez por semana
  - 4: Menos que uma vez por mês
  - 5: Não fuma esse produto
  - 9: Ignorado
  - `null`: Não aplicável
- **Frequência de Beber Álcool**: Avalia com que frequência a pessoa consome bebidas alcoólicas, categorizado de forma ordinal:
  - 1: Não bebo nunca
  - 2: Menos de uma vez por mês
  - 3: Uma vez ou mais por mês
  - 9: Ignorado
- **Quantas Doses de Álcool costuma consumir**: Este atributo mede a quantidade de doses de bebida alcoólica que o indivíduo costuma consumir em uma ocasião, variando de 1 a 99 doses.
- **Quantos dias da semana costuma ingerir álcool**: Pergunta sobre quantos dias por semana a pessoa costuma consumir álcool, com opções categóricas:
  - 1 a 7: Dias específicos da semana
  - 0: Nunca ou menos de uma vez por semana
  - 9: Ignorado
- **Sabe Ler e Escrever**: Este atributo avalia se a pessoa possui alfabetização básica, com respostas categóricas:
  - 1: Sim
  - 2: Não
- **Escolaridade**: Indica o nível de educação formal do entrevistado, sendo um valor ordinal:
  - 1: Sem Instrução
  - 2: Fundamental Incompleto
  - 3: Fundamental Completo
  - 4: Médio Incompleto
  - 5: Médio Completo
  - 6: Superior Incompleto
  - 7: Superior Completo
- **Você tem Plano de Saúde?**: Informa se o indivíduo possui algum tipo de plano de saúde médico. As respostas possíveis são:
  - 1: Sim
  - 2: Não
  - 9: Ignorado
- **Ficou Internado?**: Esse atributo informa se a pessoa foi internada por mais de 24 horas nos últimos 12 meses, categorizado como:
  - 1: Sim
  - 2: Não
  - 9: Ignorado
- **Qual foi a última vez que você foi ao médico?**: Pergunta sobre o tempo transcorrido desde a última visita do indivíduo a um médico, com categorias ordinais:
  - 1: Até 15 dias
  - 2: Mais de 15 dias até 1 mês
  - 3: Mais de 1 mês até 6 meses
  - 4: Mais de 6 meses até 1 ano
  - 5: Há mais de 1 ano
  - 9: Ignorado
- **Qual a sua percepção de estado de saúde?**: Este atributo avalia a percepção do próprio entrevistado sobre seu estado de saúde, categorizado de forma ordinal:
  - 1: Muito Boa
  - 2: Boa
  - 3: Regular
  - 4: Ruim
  - 5: Muito Ruim
  - 9: Ignorado

2.3 Etapas de pré-processamento

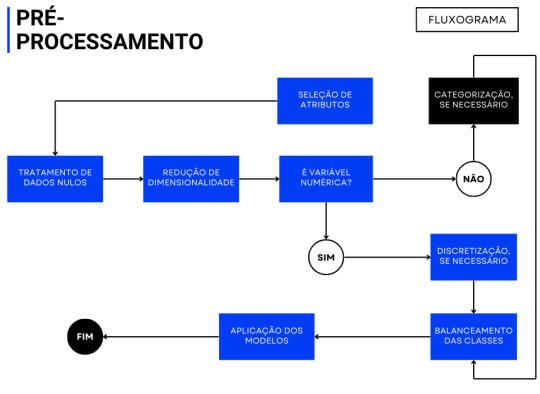

Figure 2: Fluxograma Pré-processamento

Para o pré-processamento dos dados, a figura 2 ilustra o fluxograma das etapas seguidas para melhorar a qualidade dos dados até a aplicação dos modelos.

1. **Seleção de Atributos:** Na primeira etapa, foi selecionado os atributos da PNS 2019 para a base de dados a ser utilizada futuramente pelos modelos de AM: [P00104, P00404, C006, C008, C009, P02601, P034, P050, P052, P05401, P05404, P05407, P05410, P05410, P05413, P05416, P05419, P027, P02801, P029, J037, H001, N001, I00102, D001, VDD004A, VDF004]. A soma desses atributos totaliza-se em 27.

2. **Tratando Nulos:** Nessa etapa, houve valores nulos em alguns atributos da base de dados selecionada e para cada um desses atributos foi realizado um tratamento personalizado:

- P00404 (Altura): 80.468 nulos, todos pertencendo à classe majoritária dos Saudáveis, as instâncias foram removidas.
- P00104 (Peso): 80.468 nulos, todos pertencendo à classe majoritária dos Saudáveis, as instâncias foram removidas.
- P052 (Fumou no Passado): 7.119 nulos, desses dados nulos que representam "não aplicável", 6.514 instâncias desses nulos são pessoas que responderam que fumam diariamente e por esse motivo a pergunta de fumar no passado não foi aplicada. Os outros 605 registros nulos são das pessoas que responderam que fumam menos que diariamente no passado e mais uma vez a pergunta de fumar no passado não foi aplicada. Para os 6.514 nulos, receberam o valor 1 que significa que fumaram diariamente no passado. Já os nulos restantes receberam o valor 2 que significa que fumaram no passado menos que diariamente.
- P05401, P05404, P05407, P05410, P05413, P05416, P05419 (produtos do tabaco): 47.593 nulos, todos os nulos representam as instâncias de pessoas que falaram que não são fumantes. Portanto, todos os nulos foram substituidos por 0, significando que não fumam nada desses produtos de tabaco.
- VDF004 (faixa salarial): Há 10 instâncias nulas que pertencem à classe majoritária das pessoas saudáveis e 3 pertencendo aos hipertensos, nesse caso, foram removidos.
- Quantidade de dias que o indivíduo bebe na semana: 33.702 nulos, correspondem às pessoas que responderam que não bebem nenhuma bebida álcoolica, portanto, recebeu o valor 0 sinalizando que não bebem nunca na semana.
- Quantidade de doses de álcool ingeridas: quantidade de doses que uma pessoa ingere no ato de beber, há 33.702 nulos que correspondem às mesmas instâncias do item anterior, receberam o valor 0 que significa que não ingerem nenhuma dose.

3. **Redução de Dimensionalidade:** Todas as colunas relacionadas ao Tabagismo foram fundidas em um atributo único que categoriza as pessoas de acordo com o seu nível de consumo de tabaco: Fuma Muito, Fuma Razoavelmente, Fuma Pouco e Não Fuma.

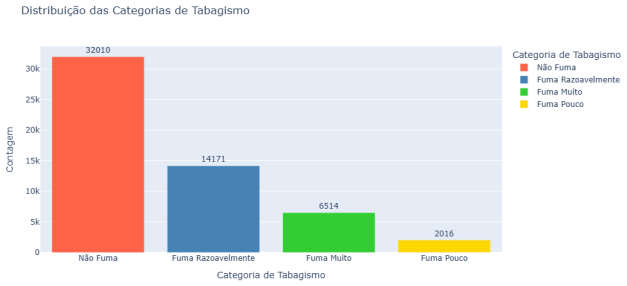

Figure 3: Tabagismo Categorizado

4. **Codificação:** Com o peso e a altura dos participantes da pesquisa de saúde, foi criada mais dois atributos: IMC e posteriormente o atributo Categoria IMC. O primeiro é baseado no cálculo do IMC descrito com a seguinte fórmula:

$$\text{IMC} = \frac{\text{Peso (kg)}}{\text{Altura (m)}^2}$$

Onde:

- **Peso (kg)**: Peso do indivíduo em quilogramas.
- **Altura (m)**: Altura do indivíduo em metros.

A categoria IMC foi codificada de acordo com as faixas de IMC:

- < 18.5 = Baixo Peso
- 18.5 - 24.9 = Peso Ideal
- 24.9 - 29.9 = Sobrepeso
- >= 29.9 = Obeso

Ademais, a genética é um fator de risco para desenvolvimento de HA e DCV, principalmente para pessoas pardas e pretas que possuem uma preponderância. Diante de tal fato, o atributo C009 foi binarizado em Pretos e Não Pretos, pessoas brancas, amarelas, indigenas e os que ignoraram a pergunta foram agrupados. Assim, o resultado ficou 33.997 pessoas pretas e 20.715 pessoas não pretas.

4. **Balanceamento das Classes:** Para o balanceamento das classes, foi preciso realizar o Undersampling das classes Saudável e Hipertenso para manter a mesma quantidade de instâncias em relação à classe das pessoas com HA e DCV. Primeiramente, o conjunto de dados foi dividido em treino e teste, sendo o conjunto de teste 20% do conjunto total, logo, o conjunto de treino teve as quantidades de pessoas com HA, saudáveis e pessoas com HA e DCV no valor de 2.342

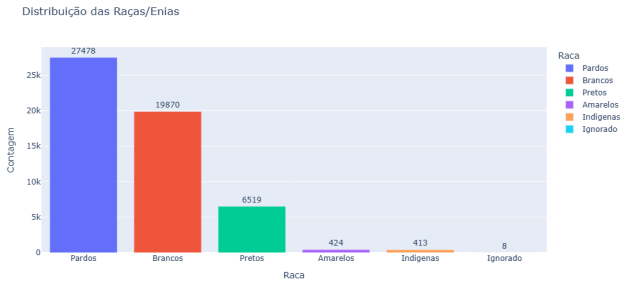

Figure 4: Raças/Etnias

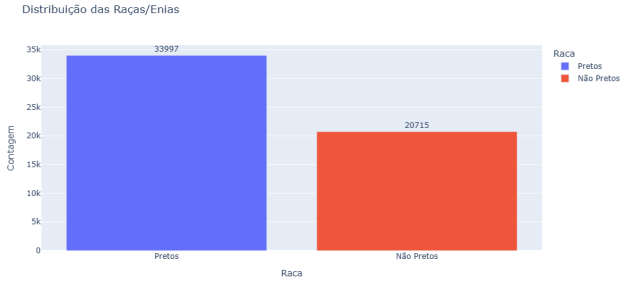

Figure 5: Raças/Etnias binarizadas

instâncias por classe. Já o conjunto de teste: HA 3.941, HA e DCV 586 e Saudáveis 6.264 instâncias.

2.4 Métricas de avaliação de qualidade

O modelo de AM foi avaliado por meio das seguintes métricas de avaliação:

- Precision: Consiste na proporção de verdadeiros positivos entre os casos que foram classificados como positivos.

$$\text{Precision} = \frac{TP}{TP + FP}$$

- Recall: Consiste na proporção de verdadeiros positivos entre todos os casos que são de fato, positivos.

$$\text{Recall} = \frac{TP}{TP + FN}$$

- F1-Score: É a média harmônica entre o Precision e o Recall.

$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

3 RESULTADOS

O modelo XGBoost trouxe o melhor resultado com a maior acurácia (64.77%), logo em seguida há o RandomForest (64.23%)

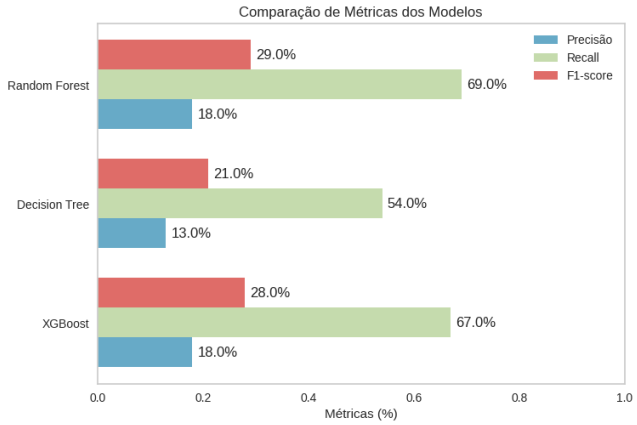

Figure 6: Métricas de avaliação dos modelos

e a Árvore de Decisão (55.6%). Isso significa que, o algoritmo XGBoost teve o maior número de instâncias classificadas como corretas em relação ao total de instâncias do conjunto de teste. À respeito da métrica de Recall, Precision e F1-Score para cada algoritmo está presente a seguir:

Table 1: Desempenho do Modelo XGBoost em Relação às Classes

	Precision	Recall	F1-Score	Instâncias
Hipertensão + Doenças Cardiovasculares	18%	67%	28%	586
Hipertensão	60%	47%	53%	3941
Saudáveis	86%	81%	81%	6264

Table 2: Desempenho do Modelo Random Forest em Relação às Classes

	Precision	Recall	F1-Score	Instâncias
Hipertensão + Doenças Cardiovasculares	18%	69%	29%	586
Hipertensão	59%	44%	51%	3941
Saudáveis	85%	76%	67%	6264

Table 3: Desempenho do Modelo de Árvore de Decisão em Relação às Classes

	Precision	Recall	F1-Score	Instâncias
Hipertensão + Doenças Cardiovasculares	13%	54%	21%	586
Hipertensão	50%	43%	46%	3941
Saudáveis	81%	64%	71%	6264

Com a análise das tabelas, fica evidente que o melhor desempenho é do XGBoost, principalmente aplicado ao contexto do problema que é o campo da saúde. Nesse campo, a prioridade é potencializar a métrica de avaliação Recall porque ela minimiza a quantidade de falsos negativos. Em outras palavras, um paciente receber um diagnóstico incorreto de saúde (ou seja, ser classificado como saudável quando na verdade apresenta uma condição de hipertensão

ou doenças cardiovasculares) pode ter consequências graves, incluindo a falta de tratamento adequado e um aumento nos riscos de complicações.O Recall, que mede a proporção de verdadeiros positivos em relação ao total de casos positivos reais, é essencial para garantir que a maioria dos pacientes que realmente precisam de intervenção médica sejam identificados. O XGBoost apresentou um Recall de 67% para a classe "Hipertensão + Doenças Cardiovasculares", superando os outros modelos. Isso indica que, embora possa ser melhor, esse modelo é mais eficaz em detectar corretamente os pacientes com essas condições críticas.

Por outro lado, as métricas de precisão (Precision) para a classe "Hipertensão + Doenças Cardiovasculares" foram relativamente baixas em todos os algoritmos. Isso significa que o modelo está gerando uma quantidade significativa de falsos positivos para essa classe, ou seja, está classificando muitas pessoas como tendo hipertensão e doenças cardiovasculares quando, na realidade, elas não possuem essas condições. Mais uma vez, os falsos positivos para o contexto de doenças são menos prejudiciais, por mais que possam causar ansiedade e más sensações no indivíduo que recebe o diagnóstico falso para a doença, não causam complicações médicas graves em uma situação contrária. Esse resultado pode ser um incômodo no contexto médico, porque leva a um número excessivo de diagnósticos incorretos, resultando em tratamentos desnecessários, estresse para os pacientes e sobrecarga do sistema de saúde.

Esses baixos valores de precisão mostram que os modelos estão enfrentando dificuldades em distinguir com precisão os casos de hipertensão com doenças cardiovasculares das outras classes. Isso pode ser devido à similaridade nos padrões de dados entre as condições de saúde, levando a uma confusão no processo de classificação.

4 CONSIDERAÇÕES FINAIS

Para melhorar a precisão, é importante continuar tentando ajustar o modelo e otimizando os hiperparâmetros. Outra alternativa seria melhorar a qualidade ou a quantidade dos dados de treinamento, especialmente para a classe de pessoas com HA e DCV, ou até mesmo testar novas abordagens de modelagem, como métodos de ensemble mais sofisticados ou algoritmos de deep learning.

5 UTILIZAÇÃO DE INTELIGÊNCIA ARTIFICIAL

Foi utilizada a ajuda do chat GPT da OpenAI para ajustar as tabelas e figuras na sintaxe LaTeX. Somado à isso, houve a utilização do Gemini do Google para a criação das fórmulas das métricas de classificação e das codificações realizadas na etapa de pré-processamento em formato LaTeX também.

6 CÓDIGO DESENVOLVIDO

Notebooks GitHub: Gustavo Costa