

**PONTIFÍCIA UNIDADE CATÓLICA DE MINAS GERAIS**

**Pós-Graduação em Ciência de Dados e Big Data**

**Modelagem e Preparação de Dados para Aprendizado de Máquina**

PREDIÇÃO DE HIPERCOLESTEROLEMIA EM PACIENTES BASEADA EM DADOS DA PESQUISA NACIONAL DE SAÚDE 2013

Amaranta Irania Lino

Caio Augusto de Oliveira Silva

Nayla Santos Guimarães

Pablo Rodrigues de Paula

Romero Igor Conceição de Santana

Brasil

2022

Desenho em preto e branco

Descrição gerada automaticamente com confiança média

PONTIFÍCIA UNIDADE CATÓLICA DE MINAS GERAIS

Pós-Graduação em Ciência de Dados e Big Data

Modelagem e Preparação de Dados para Aprendizado de Máquina

**PREDIÇÃO DE HIPERCOLESTEROLEMIA EM PACIENTES BASEADA EM DADOS DA PESQUISA NACIONAL DE SAÚDE 2013**

Trabalho apresentado ao prof. Luis Enrique Zárate como componente avaliativo da disciplina de Modelagem e Preparação de Dados para Aprendizado de Máquina.

Amaranta Irania Lino

Caio Augusto de Oliveira Silva

Nayla Santos Guimarães

Pablo Rodrigues de Paula

Romero Igor Conceição de Santana

Brasil

2022

**SUMÁRIO**

1. Introdução 2

2. Revisão bibliográfica 2

3. Tratamentos no modelo recebido 4

3.1. Solução para valores não tratados pela categorização da coluna IMC 4

3.2. Revisão da categorização da coluna idade 5

4. Seleção de novas variáveis para o modelo 6

5. Resultados obtidos 9

# Introdução

O presente relatório apresenta a modelagem preditiva de hipercolesterolemia a partir da base de dados da Pesquisa Nacional de Saúde (PNS) realizada pelo Instituto Brasileiro de Geografia e Estatística (IBGE) em 2013. O trabalho foi desenvolvido a partir do melhoramento de modelo previamente disponibilizado para uso no software Knime®.

Para o melhoramento, foi realizada pesquisa bibliográfica acerca da hipercolesterolemia, com identificação de causas e fatores relacionados a essa condição, seguida de desenvolvimento de mapa conceitual sobre o tema. Foi realizada então a seleção das variáveis a serem acrescentadas ao modelo previamente fornecido e revisão do tratamento feito sobre as variáveis que já se encontravam no modelo (fusão de variáveis, categorização das variáveis quantitativas etc).

Os resultados obtidos são apresentados ao final deste trabalho.

# Revisão bibliográfica

De acordo com o *Global Burden Disease 2019* e as bases de dados do Sistema Único de Saúde (SUS) até 2021, desde 1990 as doenças cardíacas são a principal causa de morte no Brasil. Dentre elas, as doenças arteriais coronarianas, isto é, as doenças causadas por obstrução arterial, são as mais representativas.

Já os acidentes vasculares cerebrais, que também podem ser causados por obstrução devido à deposição de gordura nos vasos sanguíneos, têm sido a segunda maior causa de morte de brasileiros desde 1990 (Oliveira et al., 2021).

As principais causas relacionadas ao colesterol alto são (Xavier et al., 2015):

1. Características fisiológicas e morfológicas:
   1. idade – indivíduos mais velhos tendem a apresentar maiores níveis de colesterol sérico;
   2. relação entre circunferências do abdome e do quadril¹;
   3. índice de massa corpórea (IMC);
   4. período pós menopausa/climatério para mulheres² – a redução dos níveis de estrogênio devido à menopausa favorece, entre outros fatores, o aumento da gordura corporal e os níveis séricos de colesterol (Conte e Franz, 2015);
   5. raça³.
2. Estilo de vida:
   1. alimentação com baixo consumo de frutas, verduras, cereais, leguminosas e peixes e alto consumo de carnes vermelhas, gorduras, sucos artificiais e açúcares (Conte e Franz, 2015);
   2. tabagismo;
   3. consumo elevado de bebidas alcóolicas;
   4. sedentarismo.
3. Comorbidades e histórico familiar:
   * + 1. diabetes melito tipos 1 e 2;
       2. doença renal crônica;
       3. doença hepática;
       4. doença autoimune, como artrite reumatoide e lúpus;
       5. histórico de hipercolesterolemia familiar4.

A partir da revisão bibliográfica, foi construído o mapa conceitual para as causas relacionadas à hipercolesterolemia (vide Figura 1, p. 3).

Figura 1. Mapa Conceitual: Hipercolesterolemia.



# Tratamentos no modelo recebido

## Solução para valores não tratados pela categorização da coluna IMC

Inicialmente, observou-se que o resultado final de acurácia estabelecido no modelo vigente, era de 65%, com um *F-measure* de 66,6% para casos de assertividade positiva e 63,2% de assertividade em casos negativos.

Figura 2. Acurácia do modelo fornecido.

Tabela

Descrição gerada automaticamente

Notou-se, na análise do modelo fornecido para trabalho, que as regras imputadas para categorizar os valores de índice de massa corporal (IMC) por faixa deixava 288 campos vazios na coluna IMC, em razão da fórmula aplicada na “*Rule Engine*” desconsiderar os valores entre um limite intervalar e outro definido na regra, a saber, os intervalos:

* (24,9 – 25);
* (29,9 – 30);
* (34,9 – 35);
* (39,9 – 40).

Figura 3. Resultado da *rule engine* para a coluna IMC no projeto fornecido: dados ausentes.

Tabela

Descrição gerada automaticamente

Dessa forma, o nó “*Rule Engine*” IMC foi ajustado para incluir também os valores pertencentes aos intervalos supracitados.

* *Rule engine* fornecida:

$IMC$ < 18.5 => "Magreza"

$IMC$ >= 18.5 AND $IMC$ <=24.9 => "Saudável"

$IMC$ >= 25.0 AND $IMC$ <=29.9 => "Sobre-Peso"

$IMC$ >= 30.0 AND $IMC$ <=34.9 => "Obesidade 1"

$IMC$ >= 35.0 AND $IMC$ <=39.9 => "Obesidade 2"

$IMC$ > 40 => "Obesidade 3"

* *Rule engine* alterada:

$IMC$ < 18.5 => "Magreza"

$IMC$ >= 18.5 AND $IMC$ < 25 => "Saudável"

$IMC$ >= 25.0 AND $IMC$ < 30 => "Sobre-Peso"

$IMC$ >= 30.0 AND $IMC$ < 35 => "Obesidade 1"

$IMC$ >= 35.0 AND $IMC$ < 40 => "Obesidade 2"

$IMC$ >= 40 => "Obesidade 3"

A alteração resultou na inclusão de mais 168 campos de IMC à base de dados, restando ainda 120 ausentes, uma vez que nestes a circunferência abdominal não constava da base, impedindo o cálculo do IMC dos mesmos. Em seguida, foram então excluídos os 120 registros de IMC com dados ausentes, por se tratar de uma representatividade irrisória.

Outra alteração de mesma ordem, foi o restabelecimento de registros ausentes que haviam sido filtradas em análises anteriores de “Comida Saudável”.

No Node “Row Filter – Del: Aus: Comid-Sau” haviam sido filtrados 1985 registros que por haver elementos ausentes foram excluídos permanentemente do aprendizado de máquina. Ocorreu que a variável “ComidaSaudavel” acabou não sendo utilizada ao final de todas as análises, entretanto, ainda assim, os 1985 anteriormente dados filtrados continuavam sendo desconsiderados, o que seria prejudicial de modo geral ao modelo caso assim permanecesse, então todos eles foram restabelecidos pela exclusão do referido Node.

Estas alterações corretivas não surtiram grande efeito de imediato ao resultado final, permanecendo a acurácia em torno ainda de 65%, mas acreditava-se, como será visto mais a frente, que o aproveitamento de mais campos traria melhores resultados a outras alterações futuras.

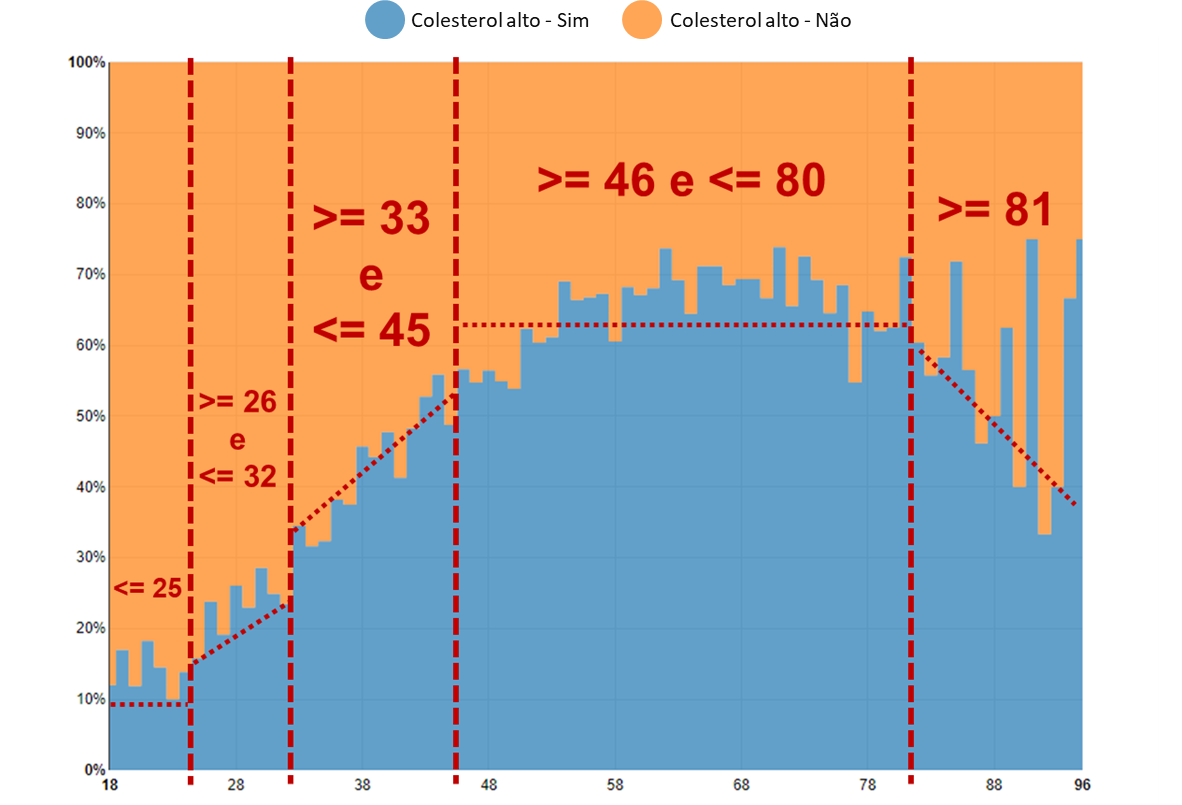
## Revisão da categorização da coluna idade

A literatura médica aponta para maior risco cardíaco nos homens a partir dos 45 anos e nas mulheres a partir dos 55 anos, quando tipicamente ocorre a menopausa (Varella, 2011). Da avaliação entre a relação entre idade e o colesterol alto (vide Figura 4), observou maior correlação positiva entre os adultos acima dos 45 anos, resultado aderente àquele preconizado pela literatura médica. Por esse motivo, optou-se por estabelecer uma categoria que iniciasse com 46 anos de idade.

Por outro lado, analisando-se os dados disponíveis e o comportamento da curva divisória entre as áreas de diagnóstico positivo e negativo para o colesterol alto, optou-se também por ajustar os limites etários das categorias, totalizando 5 categorias de idade: “Jovem”, “Adulto 1”, “Adulto 2”, “Adulto 3” e “Idoso”.

A segmentação da idade em cinco categorias, além de conveniente pela inspeção visual e literatura médica, foi a que trouxe maior acréscimo na acurácia do modelo. Considerou-se, inclusive, a divisão da faixa dos 46 aos 80 anos em outros dois segmentos, porém houve queda na acurácia; foi mantida, assim, a segmentação apresentada na Figura 4.

Figura 4. Segmentação realizada por faixa etária de acordo com o diagnóstico de colesterol alto.



Portanto, a regra ficou estabelecida assim após as reavaliações do Node "Categorização Idade" (vide Figura 5).

Figura 5. Regra para categorização da idade em cinco faixas etárias.

Texto

Descrição gerada automaticamente

O teste do modelo apresentou acurácia de 66,9%. O ganho apresentou razoável balanceamento entre os *F-measure* “Sim” (68,3%) e “Não” (65,5%) (vide Figura 6).

Figura 6. Estatísticas de acurácia após nova segmentação das faixas etárias.

Tabela

Descrição gerada automaticamente

# Seleção de novas variáveis para o modelo

Após os ajustes realizados na seção anterior, com base no mapa conceitual (Figura 1), foi realizada análise exploratória da relação entre os itens de questionário da PNS 2013 referentes às causas elencadas no mapa e o diagnóstico de colesterol alto dos entrevistados, para verificar se apresentaria o comportamento descrito na literatura médica – isto é, se haveria mais ocorrências do diagnóstico de colesterol alto em casos positivos da condição avaliada.

Observou-se que a utilização dos itens referentes ao diagnóstico prévio de doenças seria conveniente porque as respostas a esses itens, em específico, foram cuidadosamente e integralmente coletadas na pesquisa IBGE e não havia nenhum dado vazio ou ausente nessas variáveis.

Foi encontrada correlação positiva entre o diagnóstico positivo para as doenças e o elevado nível de colesterol sérico para os indivíduos nos seguintes casos (vide na página a seguir a análise gráfica):

1. Doença crônica no núcleo familiar – J007;
2. Pressão alta – Q002;
3. Diabetes – Q030;
4. Doenças cardíacas – Q063;
5. AVC ou derrame cerebral – Q068;
6. Reumatismo ou artrite – Q079;
7. Depressão – Q092;
8. Outras doenças mentais – Q110;
9. Doenças pulmonares – Q116;
10. Câncer – Q120;
11. Insuficiência renal – Q124;
12. Outras doenças crônicas – Q128.

Motivados pelos estudos científicos de especialidade médica (Brenner, 2010; Carla Mercado, 2015; e Williams, 2007), compreendeu-se que há uma estreita relação positiva entre casos de OUTRAS DOENÇAS com a existência de HIPERCOLESTEROLEMIA, especialmente diabetes e hipertensão. Analisou-se, portanto, que alguma inclusão dessas variáveis de doenças acima listadas implicaria numa melhoria do modelo da árvore de aprendizado.

**O modelo com melhor acurácia foi aquele onde se (1) incluiu a variável Q002 (diagnóstico de HIPERTENSÃO do entrevistado) e (2) uma coluna de fusão das variáveis J007, Q030, Q063, Q079, Q092, Q124 e Q128 na lista de variáveis da árvore de decisão**. Na fusão dessas variáveis, os indivíduos foram categorizados como “TemDoença”, se a resposta a qualquer um dos itens fosse “Sim” (vide Figura 7). Enquanto foram categorizados como “NãoTemDoença” caso não esteja presente nenhuma das outras doenças. **A acurácia média do modelo nas simulações realizadas foi de 69,6%, tendo sido o menor resultado observado nas simulações igual a 67,6% e o maior, 71%.**

Figura 7. Fusão das variáveis para criação da coluna “TemDoença”.

Interface gráfica do usuário, Texto, Aplicativo

Descrição gerada automaticamente

A utilização das variáveis Q068, Q110, Q116 e Q120 não resultou em melhoria da acurácia, e por isso não foram utilizadas no modelo final.

**Análise da correlação entre itens da PNS 2013 correspondentes a doenças crônicas e o diagnóstico de colesterol alto –** O elevado percentual de casos de colesterol alto entre os portadores de doença (proporção ≥ 65%) frente à proporção de 41% a 50% de casos de colesterol alto entre não portadores das doenças foi tomado como evidência da correlação positiva entre as doenças pesquisadas e o diagnóstico em colesterol alto.





# Resultados obtidos

**Conforme discutido na seção anterior, a acurácia média obtida com o modelo implementado foi de 69,6%.**

**Especificamente no experimento selecionado para evidências a seguir, o F-measure de verdadeiros positivos foi igual a 72% e de verdadeiros negativos foi de 66,6%**, o que significa que o experimento teve 72% de assertividade ao identificar um indivíduo como portador de colesterol alto, porém 5,4% a menos de probabilidade de identificar acertadamente que um indivíduo não o possui (vide Figura 8). Foram utilizados para treino 80% da amostra disponível após limpeza de registros com valores nulos de IMC e hipertensão, sendo os outros 20% empregados pra teste da acurácia do modelo (vide Figura 9).

Figura 8. Estatísticas de acurácia do modelo final.

Interface gráfica do usuário, Tabela

Descrição gerada automaticamente

Figura 9. Matriz de confusão do modelo final.

Interface gráfica do usuário, Tabela

Descrição gerada automaticamente

Ressalte-se que as evidências acima apresentadas são medidas probabilísticas, podendo variar minimamente para menos (~69%) e para mais (~70,6%), pois apresentam uma variação a cada experimento randômico, uma vez que o Node “Partitioning”, configurado como “80/20” e “Draw randownly”, apresenta esta característica.

Inclusive, foram observados outros resultados mais balanceados de Erro Alfa e Erro Beta, em torno de ~70%.

Sendo assim, realizados inúmeras outras análises, avaliações e estudos, não foram encontrados outros meios de melhoria do modelo de aprendizado de máquina, que pudessem ultrapassar uma acurácia de ~70%.

Resta ainda comentar para futuros estudos e modelos que caso tivessem sido coletados na pesquisa os dados de medida do quadril, a relação entre quadril e circunferência abdominal traria grande potencial de melhora ao modelo, já que também foi uma métrica consideravelmente relatada pela literatura médica.

**Referências bibliográficas**

Brenner et al., **Comparison of body mass index and waist circumference as predictors of cardiometabolic health in a population of young Canadian adults**. Diabetology & Metabolic Syndrome, 2010, 2:28.

Carla Mercado, PhD; Ariadne K. DeSimone; Erika Odom, PhD; Cathleen Gillespie, MS1; Carma Ayala, PhD; Fleetwood Loustalot, PhD. **Prevalence of Cholesterol Treatment Eligibility and Medication Use Among Adults** — United States, 2005–2012. Morbidity and Mortality Weekly Report Weekly / Vol. 64 / No. 47. December 4, 2015.

Conte F. A.; Franz L. B. B. **Estado Nutricional e de Saúde em Mulheres Pós-Menopausa. Santa Maria, Brasil, 2015.**

IBGE. **Pesquisa Nacional de Saúde – 2013. Questionário dos Moradores do Domicílio. Instituto Brasileiro de Geografia e Estatística, Rio de Janeiro, Brasil, 2013.**

IBGE. **Pesquisa Nacional de Saúde 2013: Acesso e Utilização dos Serviços de Saúde, Acidentes e Violências. Brasil, Grandes Regiões e Unidades da Federação. Instituto Brasileiro de Geografia e Estatística, Brasil, 2015.**

**Oliveira GMM, Brant LCC, Polanczyk CA, Malta DC, Biolo A, Nascimento BR, Souza MFM, Lorenzo AR, Fagundes Júnior AAP, Schaan BD, Castilho FM, Cesena FHY, Soares GP, Xavier Júnior GF, Barreto-Filho JAS, Passaglia LG, Pinto-Filho MM, Carrion JM, Bittencourt MS, Pontes Neto OM, Villela PB, Teixeira RA, Sampaio RO, Gaziano TA, Perel P, Roth GA, Ribeiro ALP. Estatística Cardiovascular – Brasil 2021. Brasil, 2021. Disponível em: https://doi.org/10.36660/abc.20211012. Acesso em 18/07/2022.**

**Sá A. C. M. G. N., Machado I. E., Bernal R. T. I., Malta D. C. Fatores Associados ao LDL-Colesterol Aumentado na População Brasileira Adulta: Pesquisa Nacional de Saúde. UFMG, Brasil, 2015.**

**Varella, D. Como manter o seu colesterol sob controle. Brasil, 2011. Disponível em: https://drauziovarella.uol.com.br/drauzio/artigos/como-manter-seu-colesterol-sob-controle-artigo/. Acesso em 30/07/2022.**

Xavier H. T., Izar M. C., Faria Neto J. R., Assad M. H., Rocha V. Z., Sposito A. C., Fonseca F. A., dos Santos J. E., Santos R. D., Bertolami M. C., Faludi A. A., Martinez T. L. R., Diament J., Guimarães A., Forti N. A., Moriguchi E., Chagas A. C. P., Coelho O. R., Ramires J. A. F. **V Diretriz Brasileira de Dislipidemias e Prevenção da Aterosclerose. Sociedade Brasileira de Cardiologia, Rio de Janeiro, Brasil, 2013.**

**Williams PT, Hoffman K, La I. **Weight-related increases in hypertension, hypercholesterolemia, and diabetes risk in normal weight male and female runners**. Arterioscleros Thromb Vasc Biol. 2007; 27:1811–1819.**