A Epidemia Silenciosa: Explorando os Determinantes Comportamentais e Socioeconômicos da Deficiência Renal Crônica no Brasil

Marco T. Sousa[1], **Luis E. Zarate**[1]

[1]Curso Ciência de Dados, Pontifícia Universidade Católica de Minas Gerais
Belo Horizonte – MG – Brasil

Abstract. *According to the National Health Survey (PNS), Chronic Kidney Disease (CKD) affects around two million people. CKD is a medical condition that results from a complex interaction of biological and social factors, and is known to have difficulties in both diagnosis and prognosis. Therefore, the objective is to apply machine learning techniques to characterize the profile of individuals with this disease considering various socioeconomic, health and other factors. This work proposes a method for the discovery of new knowledge with the application of Decision Tree, Naive Bayes, and Random Forest algorithms, in witch this latest has achieved the best performance. A discussion of the identified factors, including the strong relationship between hypertension, salt consumption, type of city where you live, etc.*

Resumo. *De acordo com a Pesquisa Nacional da Saúde (PNS), a Deficiência Renal Crônica (DRC) afeta cerca de dois milhões de pessoas. A DRC é uma condição médica que resulta de uma interação complexa de fatores biológicos e sociais, e é conhecida por ter dificuldades tanto no diagnóstico quanto no prognóstico. Diante disso, o objetivo é aplicar técnicas de aprendizado de máquina para caracterizar o perfil de indivíduos portadores da doença considerando diversos fatores socioeconômicos, de saúde, etc. O trabalho propõe um método para descoberta de conhecimento com a aplicação dos algoritmos Árvore de Decisão, Naive Bayes, e Floresta Aleatória, tendo este último alcançado o melhor desempenho. Esse estudo fornece uma discussão sobre os fatores identificados, incluindo a forte relação entre hipertensão, consumo de sal, tipo de cidade onde mora, etc.*

1. Introdução

De acordo com a Organização Mundial da Saúde (OMS) [OMS 2022], a Deficiência Renal Crônica (DRC) é uma condição médica grave que difere de problemas renais temporários e possui um impacto significativo na vida do indivíduo doente. Esta condição é um problema de saúde relativamente sério, principalmente por se tratar de um problema que pode durar por toda a vida do indivíduo.

A DRC é uma condição que causa grande sofrimento aos afetados e interfere significativamente em seu desenvolvimento profissional, social e familiar. Para a Sociedade Brasileira de Nefrologia (SBN) [SBN 2023], aproximadamente 140 mil brasileiros sofrem de DRC, o que nos leva ao impressionante número de cerca de 700 milhões de pessoas mundialmente [OMS 2022]. Então, pesquisas que contribuam para o diagnóstico

e tratamento desta condição são essenciais, principalmente quando se trata de pacientes em estágios iniciais da doença.

Nos casos mais graves, a DRC pode levar à insuficiência renal terminal, em que o individuo doente necessita de diálise ou transplante renal para sobreviver. De acordo com a OMS, a DRC está entre as principais causas de morte prematura em todo o mundo, e a falta de acesso a tratamentos eficazes é uma barreira para uma melhor qualidade de vida dos pacientes [Souza and Lima 2022]. Fatores que corroboram para esse problema são a falta de recursos assistenciais, a falta de profissionais de saúde, e principalmente a dificuldade para diagnosticar precocemente se o paciente realmente possui a doença ou não, o que consequentemente pode causar a postergação ou interrupção de tratamentos que poderiam ser efetivos para tratar a doença.

No contexto da Ciência de Dados, especialmente na área da saúde, o processo de descoberta de conhecimento em banco de dados (Data Mining) é bastante requisitado, pois permite analisar dados acerca de um domínio de problema, numa perspectiva multifatorial, a partir de algoritmos computacionais de aprendizado. Especificamente na área de Aprendizado de Máquina (AM), existem trabalhos contribuindo para o entendimento e diagnóstico da DRC. Um exemplo é o trabalho Oliveira e Santos (2023), onde são aplicados algoritmos de AM para predição da DRC [Oliveira and Santos 2023]. Outro estudo relevante é Pereira e Silva (2022), que utiliza técnicas de aprendizado de máquina para o diagnóstico da DRC, destacando a importância de diferentes fatores clínicos na predição da doença [Pereira and Silva 2022]. Esses trabalhos evidenciam que a identificação de transtornos renais pode ser aprimorada através da observação de características que possuem diferentes níveis de importância, muitas vezes não avaliadas por exames médicos tradicionais.

Sendo assim, o presente trabalho tem por objetivo caracterizar o adulto brasileiro com DRC por meio de um processo de descoberta de conhecimento e construção de modelos de aprendizado de máquina baseado em árvore de decisão, floresta aleatória e classificador bayesiano. Para isso, será considerado o mais recente estudo do Instituto Brasileiro de Geografia e Estatística (IBGE), Pesquisa Nacional de Saúde (PNS) 2019, pesquisa realizada por meio de questionários em todo o território nacional para retratar o perfil de saúde do cidadão brasileiro.

O artigo está organizado da seguinte forma: na segunda seção, são apresentados os trabalhos relacionados ao tema do presente estudo. A terceira seção, descreve a metodologia adotada, e a quarta seção, experimentos e análise dos resultados, são apresentados. Finalmente as considerações finais e trabalhos futuros são apontados.

2. Trabalhos relacionados

Na análise da interligação da DRC com outros distúrbios clínicos, a literatura destaca a hipertensão, a diabetes mellitus, e as doenças cardiovasculares. Na referência Pereira e Silva (2022), os autores buscaram analisar pacientes com DRC, com diabetes, e sem diabetes. A presença de hipertensão e doenças cardiovasculares, assim como sintomas relacionados à albuminúria, foram significativamente associados à presença de DRC com comorbidades[Pereira and Silva 2022]. Em Souza e Lima (2022), os autores avaliaram a presença de sintomas de hipertensão em pacientes com DRC. Os resultados sugerem que um subconjunto considerável de pacientes com DRC pode ter hipertensão clínica ou estar

em alto risco de desenvolver[Souza and Lima 2022].

Nos estudos sobre características comportamentais, a literatura têm avaliado os sintomas específicos e seu impacto na qualidade de vida. Fernandes mediu a qualidade de vida de pacientes com DRC e do grupo de controle usando regressão linear múltipla [Fernandes and Costa 2021], e observou que pacientes com DRC apresentaram baixa qualidade de vida e mais sintomas comparados ao grupo controle. De acordo com os autores, sintomas de fadiga e depressão gerados pela DRC estavam correlacionados com a baixa qualidade de vida.

Finalmente, características socioambientais também foram estudadas por Oliveira e Santos (2023). Os autores analisaram o impacto de fatores socioeconômicos na presença e severidade da DRC no futuro. Os autores concluíram que a presença de condições socioeconômicas desfavoráveis está relacionada a uma maior severidade de sintomas de DRC.

Durante a revisão de literatura realizada, não foram encontrados trabalhos que abordassem a análise da DRC utilizando especificamente dados da PNS 2019 e técnicas de aprendizado de máquina. O diferencial do presente trabalho corresponde à caracterização do perfil do paciente de DRC para a população brasileira, utilizando a base de dados PNS 2019 que possui dados socioeconômicos, alimentares, de saúde, dentre outros. Neste trabalho, são comparados os desempenhos de algoritmos caixa-branca mais interpretáveis e caixa-preta tipicamente com melhor desempenho [Loyola-González 2019].

3. Material e Métodos

3.1. Conjunto de Dados

A base de dados utilizada neste trabalho, como já mencionado, corresponde aos dados da Pesquisa Nacional da Saúde (https://www.pns.icict.fiocruz.br/PNS 2019). A PNS, analisa a percepção do estado de saúde, estilos de vida, doenças crônicas e saúde bucal da população brasileira. A PNS constitui uma base sólida para análise de políticas públicas, implementadas pelo Estado Brasileiro, a partir de um retrato da saúde da população brasileira. A base de dados original da PNS-2019 possui 1,088 atributos organizados em 26 módulos, e 293,726 registros devidamente anonimizados. A pesquisa foi aprovada pela Comissão Nacional de Ética em Pesquisa – CONEP pelo parecer 3.529.376 de Agosto de 2019. Dentro do contexto da PNS 2019, foi realizado um corte de análise voltada para a DRC, aonde ocorreu uma pré-seleção conceitual de atributos que tenham relação para o diagnóstico/caracterização da doença.

3.2. Entendimento do domínio do problema

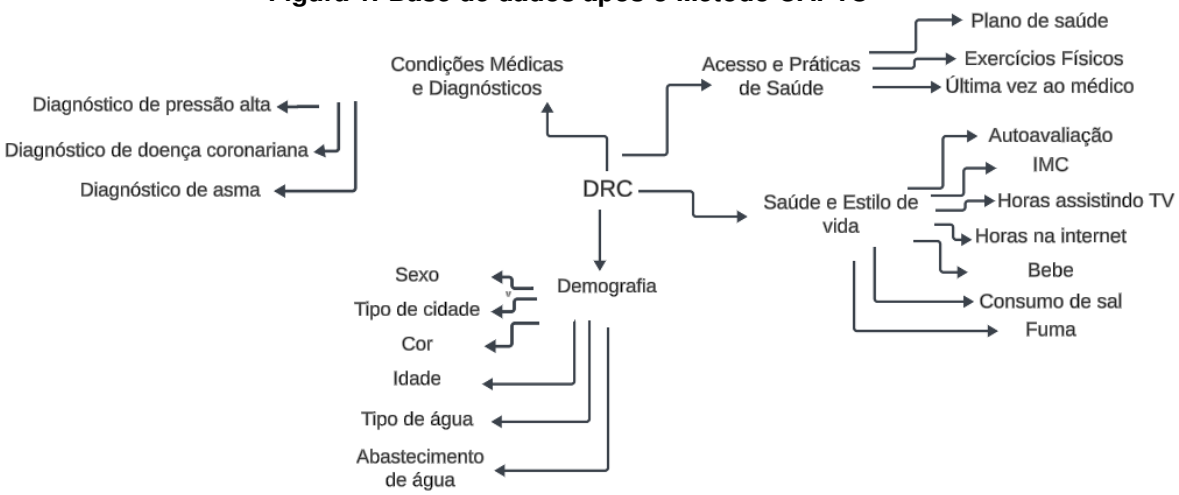
Devido à grande quantidade de atributos disponíveis na base de dados PNS, é proposta um processo para seleção conceitual dos principais atributos que podem contribuir para traçar o perfil de pessoas com DRC. Nesta etapa, é ressaltado a importância do entendimento do domínio da aplicação e da relevância do conhecimento a priori acerca desse, antes da construção de modelos de aprendizado.

Como primeiro passo do método proposto, é construído um Mapa Conceitual (MC) baseado no método CAPTO, recentemente proposto em [Zárate et al. 2023], que é uma abordagem que propõe a captura de conhecimento explícito e tácito para o entendimento de domínios do problema. O mapa conceitual captura as principais dimensões

e aspectos acerca do domínio. Após construção do MC, é realizada uma a seleção conceitual de atributos com o objetivo de obter uma base de dados mais condizente com o domínio de problema considerado.

Para construir um MC, é necessário a definição do problema a ser tratado. Neste trabalho, o problema consiste na caracterização do diagnóstico da DRC, o que demandou uma profunda revisão da literatura para adquirir conhecimento explícito. Primeiramente, foi realizada uma entrevista com especialista de domínio, da área de nefrologia, para captura do conhecimento tácito. Durante a entrevista, foi apontado as principais características clínicas e comportamentais de um indivíduo com DRC que podem ser utilizadas para o diagnóstico da doença. Foram também identificadas características socioambientais que podem influenciar na progressão dos sintomas. As discussões geraram diferentes visões sobre o tema, e as dimensões e aspectos do MC resultante podem ser vistas na Figura 1.

Figura 1. Base de dados após o Método CAPTO



A Tabela 1 mostra a relação dos atributos selecionados a partir do modelo conceitual da Figura 1. A tabela contém as Dimensões, Aspectos e Atributos da PNS, associados com a DRC. Durante esta seleção, os atributos da base de dados foram mapeados para cada uma das dimensões e, depois, subdivididos nos aspectos relevantes relacionados com a DRC.

3.3. Pré-processamento de dados

A partir dos atributos selecionados, os seguintes atributos categóricos apresentaram dados ausentes: H001 (Ultima-vez-ao-médico) e Q00201 (Diagnóstico-Hipertensão). Esses atributos tiveram seus dados faltantes imputados pela moda respectiva de cada atributo, com base no estudo da BMC (2023), no qual comprova a efetividade de substituir os dados faltantes pela moda .

De forma a reduzir a dimensionalidade, foi realizada uma combinação de atributos. Os atributos P00104, P00404 foram combinados para gerar o atributo IMC. Os atributos P027 e P029 forma combinados para gerar o atributo Frequência-consumo-álcool, e os atributos P034 e P035, combinados para gerar o atributo Frequência-fumo. A categorização do atributo IMC foi definida com base nas faixas sugeridas

Tabela 1. Descrição do Domínio DRC

Domínio: Deficiência Renal Crônica		
Dimensão: Condições Médicas e Diagnósticos		
Aspecto	Atributos Mapeados	Descrição dos Atributos
Diagnóstico: Hipertensão	Q00201	Diagnosticado com pressão alta
Diagnóstico: Asma	Q0074	Diagnosticado com asma
Diagnóstico: Doença coronariana	Q06306	Doença coronariana presente
Dimensão: Demografia		
Aspecto	Atributos Mapeados	Descrição dos Atributos
Sexo	C006	Gênero do paciente
Cor	C009	Raça/etnia do paciente
Idade	C008	Idade do paciente
Abastecimento de água	A005010	Abast. de água na casa
Tipo de água	A009010	Tipo de água ingerida
Tipo de cidade	V0026	Urbano ou rural
Dimensão: Acesso e Práticas de Saúde		
Aspecto	Atributos Mapeados	Descrição dos Atributos
Plano de saúde	I00102	Paciente tem plano de saúde
Exercícios físicos	M01601	Frequência de atividade física
Última vez ao médico	H001	Tempo da últ. consulta médica
Dimensão: Saúde e Estilo de Vida		
Aspecto	Atributos Mapeados	Descrição dos Atributos
Autoavaliação	N001	Autoavaliação de saúde
IMC	P00104, P00404	Índice de Massa Corporal
Horas TV	P04501	Tempo diário assistindo TV
Consumo sal	P02601	Nível de consumo de sal
Fuma	P034, P035	Hábitos de tabagismo
Bebe	P027, P029	Consumo de álcool
Horas Internet	P04502	Tempo diário na internet

pela OMS, e a padronização dos outros atributos foi pelas faixas sugeridas pela Secretaria da Saúde do Brasil [MS-BRASIL 2024]. Os atributos Frequência-consumo-álcool e Frequência-fumo foram codificados pelo Label Encoder, pelo grau da frequência (https://www.pns.icict.fiocruz.br/dicionário de dados da PNS).

A partir da correlação de Pearson foi realizada a análise de correlação entre os atributos. O maior valor de correlação foi 0,6 (correlação moderada), pelo qual não houve a exclusão de nenhum atributo.

De forma a confirmar a análise da Etapa 4, foi analisada a entropia no conjunto de dados. O cálculo da entropia foi realizado a partir da biblioteca *numpy* para ranquear a informação presente em cada atributo, e em relação a classe. Atributos com baixa entropia foram analisados separadamente de forma a detectar atributos com valores constantes que podem não trazer melhoramentos ao modelo de aprendizado. A análise da entropia de cada atributo com relação à classe teve como objetivo detectar e eliminar atributos com classificação direta, o que poderia levar à obtenção de modelos de aprendizado com alto desempenho, porém de pobre interpretabilidade e obviedade. Após essa análise, não foi detectado atributos para classificação direta nos modelos. Como resultado do processo de preparação de dados, o conjunto de dados resultante possui 19 atributos e 2,574 instâncias.

3.4 Modelagem

Tabela 2. Codificação por atributo

Atributo	Codificação
Q00201	1 - Diagnosticado — 2 - Não diagnosticado
Q0074	1 - Diagnosticado — 2 - Não diagnosticado
Q06306	1 - Diagnosticado — 2 - Não diagnosticado
C006	1 - Homem — 2 - Mulher
C009	1 - Branca — 2 - Preta — 3 - Amarela — 4 - Parda — 5 - Indígena
C008	Idade(Número)
A005010	(1) Rede geral — (2) Poço profundo — (3) Poço raso — (4) Fonte — (5) Água da chuva — (6) Outra
A009010	(1) Filtrada — (2) Fervida — (3) Tratada com cloro — (4) Tratada com outro — (5) Mineral — (6) Sem tratamento
V0026	(1) Urbano — (2) Rural
I00102	(1) Possui plano — (2) Não possui plano
M01601	(0) Não treina — (1) Treina de 1 a 3x por semana — (2) Treina mais de 3x por semana
H001	(1) Até 15 dias — (2) 15 dias a 1 mês — (3) 1 mês a 6 meses — (4) 6 meses a 1 ano — (5) Mais de 1 ano
N001	(1) Muito boa — (2) Boa — (3) Regular — (4) Ruim — (5) Muito ruim
P00104,P00404	IMC(Número)
P04501	(1) Menos de 1h — (2) 1 a 2h — (3) 2 a 3h — (4) 3 a 6h — (5) 6h ou mais — (6) Não usa
P02601	(1) Muito Alto — (2) Alto — (3) Moderado — (4) Baixo — (5) Muito baixo
P034,P035	(0) Não fuma — (1) Fuma de 1 a 3x por semana — (2) Fuma mais de 3x por semana
P027,P029	(0) Não bebe — (1) Bebe de 1 a 3x por semana — (2) Bebe mais de 3x por semana
P04502	(1) Menos de 1h — (2) 1 a 2h — (3) 2 a 3h — (4) 3 a 6h — (5) 6h ou mais — (6) Não usa

Tabela 3. Atributos com dados imputados

Atributo	% de Ausência
H001	65,91%
Q00201	51,78%

Os algoritmos utilizados neste estudo foram selecionados com base em sua aceitação na literatura e adequação para os objetivos propostos (interpretabilidade e desempenho, [Loyola-González 2019]). Entre os algoritmos de decisão, foi selecionado o J48 (algoritmo caixa-branca), uma implementação do algoritmo de árvore C4.5, no ambiente do JupyterLab. O segundo algoritmo selecionado foi o Random Forest (algoritmo caixa-preta, ensemble). Para o uso do algoritmo classificador Bayesiano, foi utilizado MultinomialNB. É importante ressaltar que para o classificador Bayesiano todos os atributos são considerados independentes e igualmente importantes para construção do modelo. As Tabelas 3, 4 e 5, mostram os hiperparâmetros para os algoritmos utilizados.

Tabela 4. HyP-AD

Árvore de Decisão	
Hiperparâmetro	Escolha
criterion	Gini
splitter	best
max_depth	none
min_samples_split	2
min_samples_leaf	1
max_features	none
random_state	Best

Tabela 5. HyP-NB

Naive Bayes	
Hiperparâmetro	Escolha
alpha	1
fit_prior	True
MultinomialNB	True

Tabela 6. HyP-FA

Floresta Aleatória	
Hiperparâmetro	Escolha
n_estimators	Gini
criterion	best
max_depth	None
min_samples_split	2
bootstrap	1
n_jobs	None

O conjunto de dados foi separados entre treino e teste ($hold - out$) numa proporção de 80 e 20 por cento respectivamente, distribuídos conforme tabelas abaixo: Para o processo de treinamento foi aplicado validação cruzada, com 10 dobras, e aplicado balanceamento de classes, por meio do método $undersampling$, no qual foram removidos aproximadamente 80% dos dados da classe majoritária.

Tabela 7. Separação de treino e teste

Classe	Treino	Teste
1 - Com DRC	374	173
2 - Sem DRC	1854	173

4. Resultados

Considerando os resultados de treinamento mostrados na Figura 2, o modelo baseado em Floresta Aleatória obteve o melhor desempenho para ambas as classes (1: Com DRC; 2: Sem DRC) para todas as métricas. A média para ambas as classes são: Precisão = 83%, Revocação = 83%, Acurácia = 83%, e F1-score = 83%. Para o diagnóstico clínico, é mais relevante discriminar os indivíduos com e sem doença, daí a medida F1-score (média harmônica entre a precisão e a revocação) é mais adequada para avaliar globalmente os classificadores. Os modelos interpretáveis baseados em árvore de decisão (média de F1-score = 76%) e Naive Bayes (média de F1-score = 75,7%) apresentaram desempenho similares. Além disso, como pode ser visto nas regras e na Figura 3, elas obtiveram resultados semelhantes.

Por outro lado, observando a medida de Revocação, percebe-se que o modelo interpretável, baseado em árvore de decisão, classificou adequadamente as instâncias que eram da classe "1"(Com DRC) com Revocação = 83%. Isso é muito relevante porque um bom diagnóstico para a sintomatologia elevada pode sugerir o tratamento adequado mais rapidamente. Isto significa que o profissional de saúde, família e educadores podem intervir de formas mais precoces no tratamento dos pacientes.

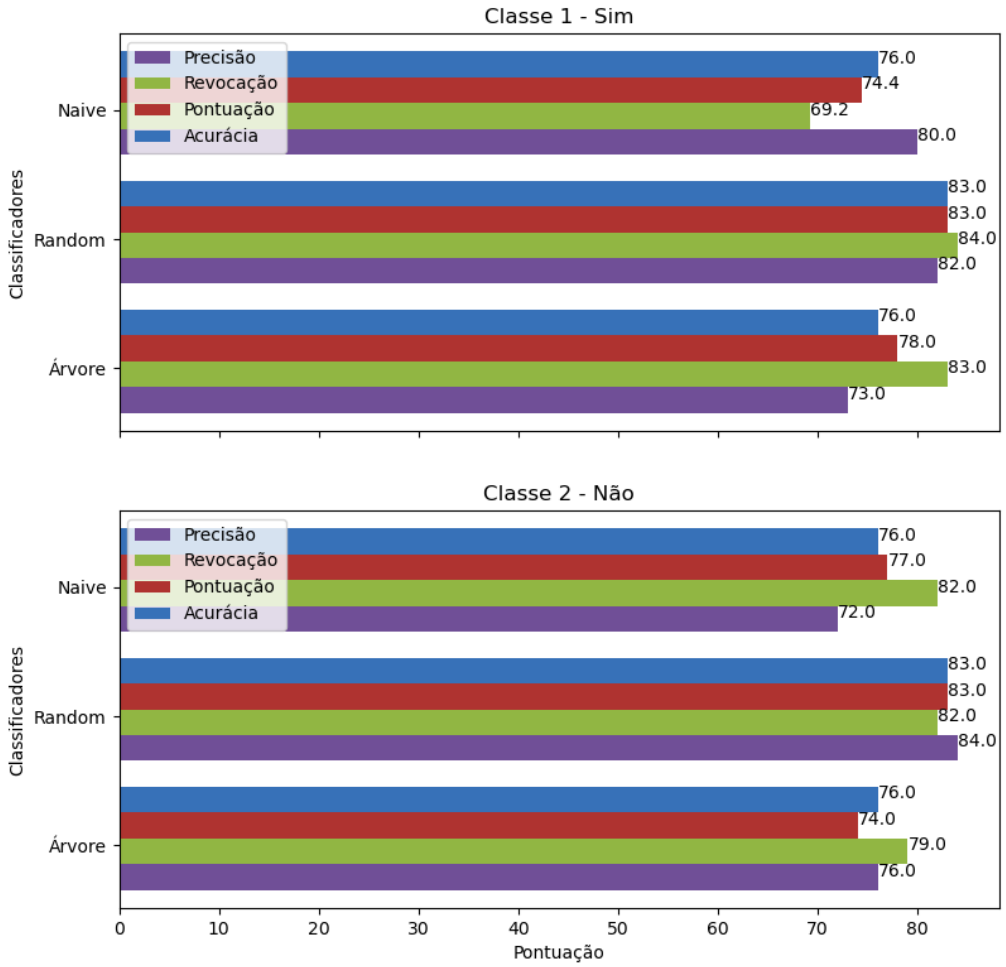
Tabela 8. Matriz de Confusão - Teste(Random Forest)

	1: Com DRC	2: Sem DRC
1: Com DRC	146	29
2: Sem DRC	28	148

Nos experimentos de teste, a partir da matriz de confusão, para o melhor modelo baseado em Floresta aleatório, e das medidas de Precisão e F-score, foi possível observar que o modelo manteve um desempenho de 83% e 82% em ambas as classes, respectivamente. A partir da Tabela 7, percebe-se que o modelo obteve um bom desempenho para o total de instâncias utilizadas.

A partir do modelo Floresta Aleatória é possível obter a relevância de cada atributo, para caracterização da DRC, apontados pelo modelo. Na Figura 3 é observado que o atributo Tipo de cidade (Urbano, Rural), é a característica mais importante, sugerindo que o tipo de cidade tem um impacto na probabilidade de um indivíduo ter DRC, ou seja, fatores como acesso a serviços de saúde, qualidade de água, e estilo de vida podem estar relacionados ao indivíduo ter a doença. Outro fator relevante é o abastecimento de água mostrando como a qualidade e o acesso a água potável podem influenciar na saúde renal. Por último,a avaliação que o indivíduo faz de sua própria saúde demostrou ser um indicador importante, mostrando que a percepção de saúde pode ser um reflexo de condições subjacentes como a DRC.

Figura 2. Comparação de resultados - base de treino

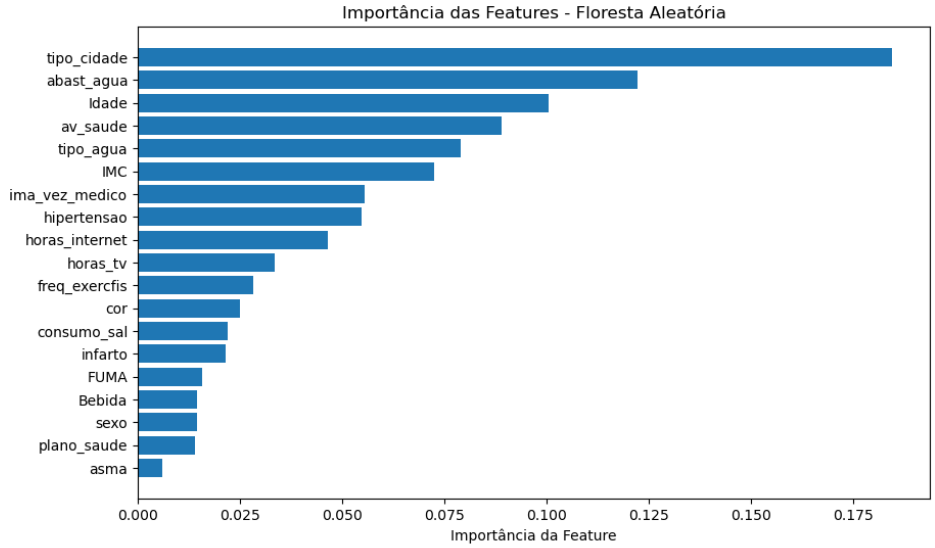


4.1. Interpretação de Regras - Árvore de decisão

Tendo a árvore de decisão alcançado um desempenho satisfatório, é aproveitado a capacidade interpretativa do modelo. Para isso as principais regras são mostradas a seguir:

Regra 1: Idade $\leq$ 47,5 and Hipertensao $\leq$ 1,5 and Ultima vez ao médico nos últimos 180 dias $\leq$ 1,5 and Tipo da cidade $\leq$ 1,5 Então Classe = Com DRC
Regra 2: Hipertensao $\leq$ 1,5 and Avaliação da saúde $\leq$ 2,5, Tipo da cidade $\leq$ 1,5 Então Classe = Com DRC
Regra 3: Idade $\leq$ 22,5 and Tipo da cidade $\leq$ 1,5 Então Classe = Sem DRC
Regra 4: Ultima vez ao médico nos últimos 180 dias $\leq$ 1,5 and Consumo de sal $\geq$ 4,5 and Asma $\leq$ 1,5 Então Classe = Com DRC
Regra 5: Hipertensao $\leq$ 1,5 and Ultima vez ao médico nos últimos 180 dias $\leq$ 1,5 and Frequência de exercícios físicos $\geq$ 2,5 Então Classe = Sem DRC

Figura 3. Feature Importance



Analisando as regras, é possível observar a recorrência de alguns atributos, como por exemplo, o diagnóstico de hipertensão. A primeira regra sugere que indivíduos com menos de 47,5 anos que têm hipertensão ($\leq$ 1,5) e que visitam o médico 1 vez em 180 dias ($\leq$ 1,5) e que vivem em cidades ($\leq$ 1,5) têm maior probabilidade de apresentar DRC. Isso pode indicar um estilo de vida com pouca atenção à saúde preventiva. Já na segunda regra, é implicado que pessoas com hipertensão ($\leq$ 1,5), com baixa percepção de saúde (avaliação de saúde $\leq$ 2,5) e que vivem em zonas urbanas ($\leq$ 1,5), também são mais propensas à DRC. A percepção negativa da própria saúde pode estar associada a uma falta de autocuidado e menor acesso a informações sobre saúde. Por fim, na quarta regra é implicado que pessoas que foram ao médico há pelo menos 1 mês ($\leq$ 1,5) que consomem pouco sal (tipo de cidade $\geq$ 4,5) e são asmáticos ($\leq$ 1,5) têm maior probabilidade de desenvolver DRC. Isso pode indicar uma vulnerabilidade dos indivíduos com asma em áreas urbanas.

5. Conclusão

O presente estudo teve como objetivo principal caracterizar a DRC na população brasileira por meio de um processo de descoberta de conhecimento e de técnicas de classificação como Árvore de Decisão, Floresta Aleatória e Naive Bayes. Embora haja uma quantidade considerável de trabalhos de aprendizado de máquina abordando DRC, não há nenhum diretamente relacionado a DRC considerando a pesquisa PNS 2019.

Analisando os resultados finais, observa-se que os modelos caixa-branca e caixa-preta (ensemble) demonstraram uma boa eficácia relevante para um diagnóstico precoce e intervenção adequada na sintomatologia. Este aspecto é importante para os profissionais de saúde, sendo possível permitir intervenções precoces que podem melhorar o tratamento dos pacientes. É importante salientar que através das técnicas de pré-processamento, podem haver limitações na distribuição dos dados, introduzindo vieses significativos. Para estudos futuros, é recomendado que possam ser obtidas análises estatísticas mais robustas,

além de análises de sensibilidade. Também podem ser usadas outra gama de classificadores (por exemplo, o grid search). Em conclusão, o uso de modelos de aprendizado de máquina, como os utilizados neste estudo, obteve uma capacidade de interpretar e identificar os fatores relevantes para a doença. A implementação de tais modelos em práticas clínicas pode melhorar a qualidade do atendimento ao paciente e possibilitar intervenções mais direcionadas e personalizadas. A análise futura pode se beneficiar da inclusão de um conjunto de dados mais abrangente e diversificado para validar ainda mais os achados e aumentar a generalização dos modelos. Além disso, a integração de outras técnicas podem proporcionar uma visão mais holística do problema, contribuindo para o avanço no campo da saúde pública e da medicina preventiva.

Agradecimentos

Os autores agradecem o apoio recebido do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Processo No 303133/2021-0, e do Fundo de Incentivo à Pesquisa (FIP) da PUC Minas, Processo No 30914-1S/2024.

Referências

Fernandes, G. and Costa, M. (2021). Machine learning techniques for early detection of chronic kidney disease in brazil. *Artificial Intelligence in Medicine*, 117(4):153–167.

Loyola-González, O. (2019). Black-box vs. white-box: Understanding their advantages and weaknesses from a practical point of view. *IEEE Access*, 7:154096–154113.

Methodology, B. M. R. (2023). Comparison of the effects of imputation methods for missing data in predictive modelling. In Example, C. and Sample, B., editors, *BMC Medical Research Methodology*, pages 15–20. BioMed Central.

MS-BRASIL (2024). Ministério de saúde, brasil. Ministério de Saúde, Brasil, gv.br.

Oliveira, A. and Santos, M. (2023). Application of machine learning algorithms in chronic kidney disease prediction: A brazilian study. *Journal of Medical Systems*, 47(2):289–305.

OMS (2022). The global impact of chronic kidney disease. In Division, W. H. S., editor, *Global Health Statistics 2023*, pages 45–60. World Health Organization.

Pereira, L. and Silva, R. (2022). Utilizing machine learning for the diagnosis of chronic kidney disease in brazilian patients. *Health Informatics Journal*, 28(1):112–125.

SBN (2023). Impacto da doença renal crônica no brasil. In Silva, J., editor, *Relatório Anual da Sociedade Brasileira de Nefrologia*, pages 23–45. Sociedade Brasileira de Nefrologia.

Souza, T. and Lima, F. (2022). Predictive models for chronic kidney disease: A study in brazilian population using machine learning. *BMC Medical Informatics and Decision Making*, 22(3):210–223.