UNIVERSIDADE FEDERAL DO MARANHÃO

CENTRO DE CIÊNCIAS EXATAS E TECNOLOGIAS

PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA DE ELETRICIDADE

RONALDO DOS SANTOS SILVA JUNIOR

Aplicação de Aprendizado de Máquina e Redes Neurais Artificiais para classificação da Doença Renal Crônica através de variáveis não invasivas

São Luís, MA

2022

RONALDO DOS SANTOS SILVA JUNIOR

Aplicação de Aprendizado de Máquina e Redes Neurais Artificiais para classificação da Doença Renal Crônica através de variáveis não invasivas

Dissertação apresentada, como requisito parcial para a obtenção do título de Mestre em Engenharia Elétrica, ao Programa de Pós-Graduação em Engenharia Elétrica da Universidade Federal do Maranhão

Orientador: Prof. Dr. Ewaldo Eder Santana Carvalho

Coorientador: Prof. Dra. Nilviane Pires Sousa

São Luís, MA

2022

Ficha gerada por meio do SIGAA/Biblioteca com dados fornecidos pelo(a) autor(a).
Diretoria Integrada de Bibliotecas/UFMA

dos Santos Silva Junior, Ronaldo.
Aplicação de Aprendizado de Máquina e Redes Neurais Artificiais para classificação da Doença Renal Crônica através de variáveis não invasivas / Ronaldo dos Santos Silva Junior. - 2022.
37 f.

Coorientador(a): Nilviane Pires Silva Sousa.
Orientador(a): Ewaldo Eder Carvalho Santana.
Dissertação (Mestrado) - Programa de Pós-graduação em Engenharia Elétrica/ccet, Universidade Federal do Maranhão, São Luís, 2022.

1. Aprendizado de máquina. 2. Doença Renal Crônica. 3. Redes Neurais Artificiais. I. Eder Carvalho Santana, Ewaldo. II. Pires Silva Sousa, Nilviane. III. Título.

Aplicação de Aprendizado de Máquina e Redes Neurais Artificiais para classificação da Doença Renal Crônica através de variáveis não invasivas

Dissertação apresentada, como requisito parcial para a obtenção do título de Mestre em Engenharia Elétrica, ao Programa de Pós-Graduação em Engenharia Elétrica da Universidade Federal do Maranhão

Aprovação em: 18/ 02 /2022

BANCA EXAMINADORA

Prof. Dr. Ewaldo Eder Carvalho Santana (Orientador)
Universidade Estadual do Maranhão – UEMA

Prof. Dr. Nilviane Pires Silva Sousa (Coorientador**)**
Universidade Federal do Maranhão – UFMA

Prof. Dr. Allan Kardec Duailibe Barros Filho (Membro Interno**)**
Universidade Federal do Maranhão – UFMA

Prof. Dra. Giselle Cutrim de Oliveira Santos (Membro Externo**)**
Universidade Estadual do Maranhão - UEMA

AGRADECIMENTOS

Agradeço primeiramente à minha mãe, Joycelene Souza, e ao meu pai, Ronaldo Silva, por todo apoio, incentivo e compreensão que me foi dado durante este percurso e que foram determinantes para essa enorme conquista.

Agradeço também à minha irmã, Giovana Silva, pelo apoio durante todos esses anos. À minha grande amiga e companheira, Cindy Lima, por seu inestimável papel nesta conquista.

Às minhas pequenas grandes amigas, Lara Beatriz, Analú Pereira e Thays Emanuelle pelo carinho e amor.

Aos meus colegas do laboratório PIB, que nunca mediram esforços para me auxiliar no que fosse necessário.

Ao Professor Carlos Magno pela oportunidade, confiança e ensinamentos que me foram dados. Ao Professor Allan Kardec pelos conselhos e conhecimentos compartilhados.

Ao Centro de Prevenção de Doenças Renais (CPDR) em especial à Dra. Erika Carneiro e à Dra. Luana Azoubel, meus sinceros agradecimentos pela disponibilidade e contribuição no processo de desenvolvimento deste trabalho.

Por fim, agradeço aos meus orientadores Dr. Ewaldo Santana e Dra. Nilviane Pires, por todo apoio e oportunidades que me foram dados durante esses anos e para o desenvolvimento deste trabalho.

RESUMO

Áreas da Inteligência Artificial (IA), o aprendizado de máquina e as redes neurais artificiais ganham destaque no desenvolvimento de técnicas, as quais exigem um preparo prévio do conjunto de dados para que o algoritmo execute a classificação automática do conjunto de dados. Nos últimos anos, técnicas de aprendizado de máquina e análise estatística vêm sendo aplicadas por pesquisadores da área da saúde, a fim de viabilizar tomada de decisões de prognósticos clínicos, gestão de assistência à saúde, diagnóstico e monitoramento de diversas doenças.A Doença Renal Crônica, uma das DCNT mais prevalentes a nível global, é caracterizada pela perda progressiva e irreversível da função renal glomerular, tubular e endócrina. .O diagnóstico precoce da DRC é considerado um grande desafio, visto que em estágios iniciais a doença é caracteristicamente assintomática e as manifestações clínicas destacam-se entre os estágios de insuficiência renal moderada a severa. A fim de realizar o diagnóstico da DRC, foram utilizados os modelos de Regressão Logística e RNA, onde foi utilizado o 5-fold crossvalidation em um conjunto de dados de 291 indivíduos maiores de 18 anos. Os modelos retornaram um bom desempenho, tendo os dois a área sobre a curva ROC (AUROC) = 0.94, e a RNA obteve uma acurácia e sensibilidade de 87%, já a Regressão obteve acurácia e sensibilidade de 85%. Dessa forma, nossos modelos obtiveram desempenho aceitável para classificar portadores da DRC, apresentando-se como alternativa de baixo custo para o rastreamento da doença.

Palavras-chave: Aprendizado de máquina; Redes Neurais Artificiais; Doença Renal Crônica.

ABSTRACT

Areas of Artificial Intelligence (AI), machine learning and artificial neural networks are underlined in the development of techniques, which require prior preparation of the database so that the algorithm performs the automatic classification of the database. In recent years, machine learning and statistical analysis techniques have been applied by health researchers in order to make clinical prognostic, decision-making, health care management, diagnosis and monitoring of various diseases feasible. Chronic Kidney Disease, one of the most globally prevalent NCDs, is characterized by the progressive and irreversible loss of glomerular, tubular and endocrine renal function. Early diagnosis of CKD is considered a great challenge, since in early stages the disease is characteristically asymptomatic and clinical manifestations stand out among the stages of moderate to severe renal failure. In order to diagnose CKD, Regression Logistic and ANN models were used, in which 5-fold cross validation was used in a dataset of 291 individuals over 18 years of age. The models had a good performance, both having the area under the ROC curve (AUROC) = 0.94, and the ANN obtained an accuracy and sensitivity of 87%, whereas the Regression obtained an accuracy and sensitivity of 85%. Thus, our models achieved acceptable performance for classifying CKD patients, presenting themselves as a low-cost alternative for the disease screening.

Key-words: Machine learning, Artificial Neural Network, Chronic Kidney Disease

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIAÇÕES

CC	Circunferência da Cintura
CQ	Circunferência do Quadril
DCNT	Doenças Crônicas Não Transmissíveis
DRC	Doença Renal Crônica
IBGE	Instituto Brasileiro Geográfico e Estatística
IMC	Índice de Massa Corporal
OMS	Organização Mundial de Saúde
PAD	Pressão Arterial Diastólica
PAS	Pressão Arterial Sistólica
RCE	Relação Cintura Estatura
SBN	Sociedade Brasileira de Nefrologia
SUS	Sistema Único de Saúde
TFG	Taxa de Filtração Glomerular
RNA	Redes Neurais Artificiais

SUMÁRIO

1. INTRODUÇÃO

Dentre as diversas áreas da Inteligência Artificial (IA), o aprendizado de máquina e as redes neurais artificiais ganham destaque no desenvolvimento de técnicas, as quais exigem um preparo prévio do conjunto de dados para que o algoritmo execute a classificação automática do conjunto de dados. As técnicas empregadas, como a mineração de dados e aprendizado de máquina, utilizam conceitos da inteligência artificial para tomar decisões baseadas em treinamentos realizados previamente (FERNANDES; FILHO, 2019).

Nos últimos anos, técnicas de aprendizado de máquina e análise estatística vêm sendo aplicadas por pesquisadores da área da saúde, a fim de viabilizar tomada de decisões de prognósticos clínicos, gestão de assistência à saúde, diagnóstico e monitoramento de diversas doenças, em especial, as Doenças Crônicas Não Transmissíveis (DCNT) (STANIFER et al., 2016). As DCNT representam um importante problema à saúde pública, uma vez que são consideradas a principal causa de morte no mundo, além de provocarem mortalidade prematura, menor qualidade de vida e uma significativa sobrecarga no sistema de saúde (WHO, 2020).

A Doença Renal Crônica, uma das DCNT mais prevalentes a nível global, é caracterizada pela perda progressiva e irreversível da função renal glomerular, tubular e endócrina. O número de casos registrados de doentes renais aumentou significativamente na última década, sendo associado a diversos fatores como envelhecimento e transição demográfica da população. (ALCALDE; KIRSZTAJN, 2018; NEVES, 2020; SCHAEFER, 2015). A classificação da DRC leva em consideração o estadiamento da doença, dividido em cinco estágios com base na Taxa de Filtração Glomerular (TFG), com o objetivo de facilitar o protocolo de tratamento e a estimativa de diagnóstico (BRASIL, 2014).

A DRC gera um forte impacto na saúde do paciente quando em estágios mais avançados, trazendo como única alternativa de tratamento a terapia renal substitutiva, além de do desenvolvimento e progressão de doenças cardiovasculares e, assim, reduzindo drasticamente a qualidade de vida do paciente e, em muitos casos, levando a óbito (PERES; BETTIN, 2015).

O diagnóstico precoce da DRC é considerado um grande desafio, visto que em estágios iniciais a doença é caracteristicamente assintomática e as manifestações clínicas destacam-se entre os estágios de insuficiência renal moderada a severa. O monitoramento da progressão da doença é realizado com base em diversos fatores clínicos, incluindo as comorbidades que a envolvem. Em pacientes diabéticos e hipertensos, a identificação precoce

da DRC representa o método mais empregado para o monitoramento contínuo desses pacientes, por meio de exames para avaliação da função renal (JÉSUS et al., 2017).

De modo geral, pacientes com DRC apresentam significativas alterações fisiológicas e bioquímicas em todos os estágios da doença, com grande associação às comorbidades de risco cardiometabólico como a obesidade, diabetes, hipertensão arterial sistêmica, dislipidemias, etc. Portanto, o monitoramento dos fatores de risco é de grande importância, especialmente, quando realizado através de métodos acessíveis e de baixo custo como a antropometria e análise do perfil bioquímico, como forma de prevenção e diagnóstico precoce (MANCINI, 2016).

1.1 Objetivos

1.1.1 Objetivos Gerais

Esta dissertação tem como objetivo geral desenvolver um classificador para identificar pacientes portadores da DRC através de variáveis não invasivas, utilizando aprendizado de máquinas e redes neurais artificiais.

1.1.2 4.2 Objetivos Específicos

- Analisar as variáveis antropométricas associadas à DRC.
- Aplicar técnicas de aprendizado de máquinas e redes neurais artificiais para classificar portadores da DRC.

1.2 Motivação

No ano de 2012, o Sistema Único de Saúde (SUS) financiou 84% do tratamento de paciente em alguma modalidade de terapia renal substitutiva, estimado em dois bilhões de reais em procedimentos ambulatoriais de hemodiálise e diálise peritonial (SILVA et al, 2016). No triênio 2013-2015, os gastos do SUS com exames para diagnóstico e tratamento da DRC totalizaram aproximadamente R$825 milhões de reais e um aumento de 10,95% (ALCALDE; KIRSZTAJN, 2018).

Através da Portaria GM/MS nº 1.675/2018, o Ministério da Saúde estabeleceu uma Rede de Atenção à Saúde (RAS) para pessoas com DRC, contemplando os mais diversos níveis de atenção, incluindo ações de prevenção e tratamento de fatores de risco, diagnóstico precoce, acesso universal e gratuito ao tratamento, medicamentos, consultas com equipe

multidisciplinar, transporte e acesso à internação hospitalar e equidade em lista de espera para transplante renal, na esfera do SUS (BRASIL, 2018).

Assim, a utilização de técnicas computacionais para o mapeamento das complicações advindas da DRC são extremamente úteis e o aprendizado de máquina e RNAs é importante para auxiliar no diagnóstico precoce e monitoramento da progressão da doença, contribuindo fortemente às estratégias já implementadas no sistema de saúde.

2 FUNDAMENTAÇÃO TEÓRICA

2.1 Aprendizado de Máquina

O aprendizado de máquina é uma das áreas mais abordadas na inteligência artificial e na classificação de dados em saúde, permitindo a identificação de padrões com base em casos e experimentos anteriores, semelhante ao que ocorre na inteligência humana. Este método engloba a criação de uma função de treinamento para um determinado conjunto de dados, utilizando um mecanismo de inferência lógica (ERICKSON et al, 2017).

O aprendizado por indução configura um tipo de inferência lógica, na qual é possível obter generalizações com base em fatos específicos, com o objetivo de induzir conceitos verdadeiros ou não. Em outras palavras, a utilização de premissas verdadeiras não garante que se chegue a conclusões verdadeiras. Este método, apesar de ser considerado uma estratégia complexa, permite o aprendizado de conceitos amplos e complexos e, por isso, é utilizado preferencialmente em estudos que envolvem o aprendizado de máquina (METZ, 2006).

Em suma, o método de aprendizagem de máquina é formado por um conjunto de procedimentos que visam adquirir informações sobre determinado domínio através de experimentos e observações. O ser humano, de modo geral, adquire conhecimento por meio da repetição de uma atividade estabelecida ou pela generalização de um fato já testemunhado. Esse processo de generalização ou reconhecimento de padrões caracteriza o aprendizado por indução, definido por um raciocínio posterior à consideração de um determinado número de casos particulares (ALPAYDIN, 2004).

O aprendizado de máquina indutivo é dividido em aprendizado supervisionado, semi-supervisionado e não-supervisionado; diferindo-se conforme os atributos de classe que rotula as observações do conjunto de dados estudado (CARBONELL; MITCHELL; MICHALSKI et al, 1983).

Na figura 1, verifica-se a hierarquia do aprendizado indutivo, no qual observamos se os dados fornecidos a um algoritmo de aprendizado de máquina podem ou não estar rotulados. Assim, são definidas duas classificações:

- **Supervisionado:** Possui um número significativo de dados rotulados.
- **Não-supervisionado:** Não possui rotulação nos dados.

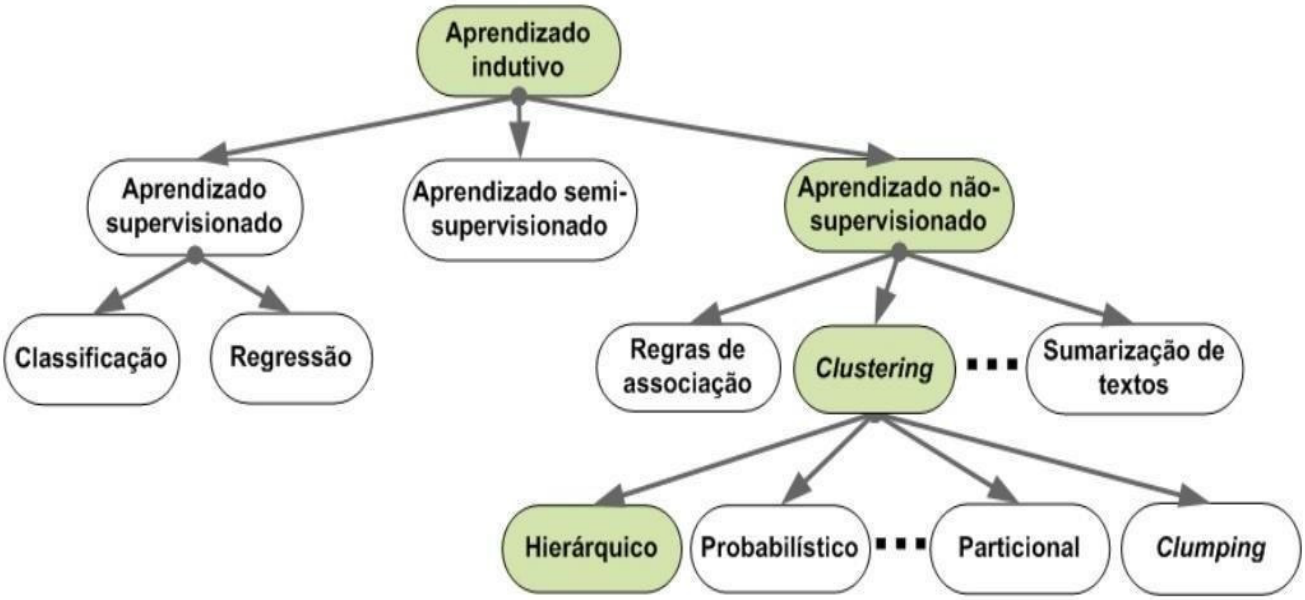

Figura 1 - Hierarquia do aprendizado indutivo (METZ, 2006)

2.2 Aprendizado supervisionado

No aprendizado de máquina supervisionado há a existência de um agente "supervisor" que já possui informações a respeito da saída, pois no processo de aprendizagem é fornecido um conjunto de registros R = {R1, R2, R3, ..., Rn}, onde para todo Ri ∈ R, tem-se um rótulo relacionado à classificação do dado Ri. Esse conjunto de dados acompanhado de seus respectivos rótulos é a origem do conhecimento que é fornecido para o algoritmo de aprendizagem para que esse desenvolva uma função *f*, que é chamada de modelo, e que a partir pode-se prever rótulos para novos registros (MONARD; BARANAUSKAS, 2003).

Neste tipo de aprendizado, a base de dados é geralmente dividida em conjuntos chamados de conjunto de treinamento e teste que são descritos a seguir (ALPAYDIN, 2004):

- Conjunto de treinamento: É o conjunto principal, pois é a partir dele que se induzem as hipóteses, e dessa forma ele deve representar a distribuição da população estudada para que a indução de classificadores possa ser feita com sucesso.
- Conjunto de teste: É utilizado para avaliar o modelo gerado. Os dados contidos neste conjunto não podem ter sido utilizados no conjunto de treinamento.

O aprendizado supervisionado pode ser caracterizado por duas funções significativas: a Classificação, que utiliza rótulos com valores discretos, que podem ser dicotômicos (0 ou 1,

doente ou não doente), ou numéricos; e a Regressão onde os rótulos são valores contínuos, valores de altura ou peso, por exemplo (FACELI et al, 2011).

2.2.1 Classificação

Um problema de classificação é identificado quando necessita-se ter conhecimento a respeito de uma categoria ou espécie de um novo caso e tanto a entrada quanto a saída são valores discretos. Por exemplo, o famoso conjunto de dados de flor *ÍRIS*, que é um conjunto de dados multivariado criado pelo estatístico e biólogo Ronald Fisher, onde são apresentados dados referentes à largura e ao comprimento das sépalas e pétalas de três espécies de flores da íris (setosa, virginica e versicolar) e a partir dos dados de uma nova flor não identificada devemos classificá-la dentro destas três espécies possíveis. Dessa forma entende-se que o objetivo na tarefa de classificação é prever a qual classe pertence um novo dado desconhecido (AMARAL, 2016).

O número de algoritmos de classificação que pode ser utilizado é bem grande, como: Algoritmo do n-Vizinho Mais Próximo, Algoritmos baseados no Teorema de Bayes como o Naive Bayes e Redes Bayesiana de Classificação, Árvores de decisões, Regressão logística, Máquina de Vetor de Suporte e as Redes Neurais Artificiais (FACELI et al, 2011).

2.2.2 Regressão Logística

Diversas funções de distribuição foram utilizadas para serem utilizadas na análise de variáveis dicotômicas, e nesse sentido, Cox (1970) e Walker e Duncan (1967) propuseram a regressão logística, que se baseia na suposição de que o logaritmo da razão das probabilidades de duas categorias é linear.

Dessa forma, seja uma variável dependente binária que indica a ausência ou não de uma determinada característica. O modelo de Regressão Logística é uma modelo particular dos Modelos Lineares Generalizados (McCaullagh e Nelder, 1989), cuja função de ligação é dada por:

$$\left\{\frac{\pi(x)Ln}{1-\pi(x)}\right\} = \beta x, \qquad (1)$$

Onde,

$\beta' = (\beta_0, \beta_1, \beta_2, \dots, \beta_p)'$: vetor de parâmetros desconhecidos;

$\pi(x)$ = E(Y = 1| x) = P(Y = 1|x): probabilidade de que o indivíduo seja portador de determinada característica, dado o vetor de variáveis x.

Esta probabilidade pode ser expressa como mostrado à seguir:

$$\pi(x) = \frac{\exp(\beta' x)}{1 + \exp(\beta' x)} \quad (2)$$

E o modelo pode ser escrito como

$$Y = \pi(x) + \varepsilon, \quad (3)$$

Onde ε é o erro aleatório do modelo.

Como Y pode assumir somente dois valores, o erro ε também pode assumir apenas dois valores, de forma que,

- se Y =1, então ε = 1 - $\pi(x)$, com probabilidade $\pi(x)$,

- se Y = 0, então ε = - $\pi(x)$, com probabilidade 1 - $\pi(x)$,

Isto é, ε segue uma distribuição com média 0 e variância $\pi(x)(1 - \pi(x))$. A distribuição convencional de Y, dado um vetor x é conhecida como distribuição de Bernoulli, com probabilidade de sucesso igual a $\pi(x)$. É possível estimar $\pi(x)$ através da estimativa de $\beta' x$. O vetor de parâmetro β é estimado utilizando-se a função de máxima verossimilhança em relação ao p + 1 elementos do vetor. Dessa forma a regra de classificação para uma população consiste em estimar $\pi(x)$ para uma nova observação x_c : se $\hat{\pi}(x_c) \geq 0{,}50$, então classificamos a observação no grupo portador de determinada característica, caso contrário, no grupo de não portadores. (Hosmer e Lemeshow, 1989).

2.2.3 Redes Neurais Artificiais

A inteligência computacional busca o desenvolvimento de sistemas que simulam os comportamentos humanos. As RNA são técnicas que simulam o sistema nervoso central do ser humano em busca de solucionar problemas. Dessa forma o conceito de aprendizado da RNA é através da experiência, ou seja, aprendendo, errando e fazendo novas descobertas.

Haykin (1994) descreve que MaCullough e Pits, em 1943, sugeriram a construção de uma modelo de máquina inspirado no cérebro humano, dando o pontapé inicial para a neurocomputação, baseado em modelos matemáticos. Em 1951 o primeiro neurocomputador denominado Snark foi construído. Snark ajustava automaticamente seus pesos entre as sinapses, mas não executava nenhuma função útil, porém serviu de inspiração para as ideias

de estruturas que o sucederam. Frank Rosenblatt concebeu o “Perceptron” em 1957. Seu interesse inicial era o reconhecimento de padrões. As pesquisas de desenvolvimento de RNAs teve seu principal ponto de partida em 1986, com a publicação do livro Parallel Distributed Processing (Processamento Distribuído Paralelo), editado por David Rumelhart e James McClelland. Em 1987, universidades anunciaram a formação de institutos de pesquisa e programas de educação em neurocomputação.

Dessa forma, o modelo de neurônio artificial proposto é baseado no modelo biológico da figura 2, onde podemos observar o corpo celular, dendritos e axônios e sua comunicação à sinapse. A comunicação entre os neurônios ocorre por meio dos sinais captados pelos dendritos que através do axônio se transporta para o dendrito de outra célula através da sinapse. Abaixo temos a representação dessa estrutura (BITTENCOURT, 2016).

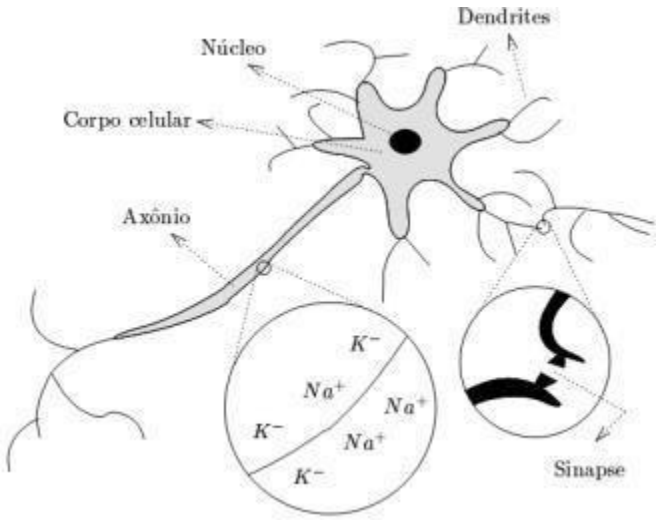

Figura 2 : Modelo de neurônio Artificial, Bittencourt (2016)

O primeiro modelo matemático para a RNA descrito por MaCullough e Pits, se tratava apenas de um dispositivo binário, como vemos na figura 3, sendo assim a saída formada por um sinal ou não sinal (ativo ou não ativo), e as entradas possuíam um ganho arbitrário. A saída era calculada por meio da soma ponderada das entradas com seus pesos como fatores de ponderação, se o resultado atingisse certo valor, a saída era ativada, caso contrário, não (ZAMBIASI, 2002).

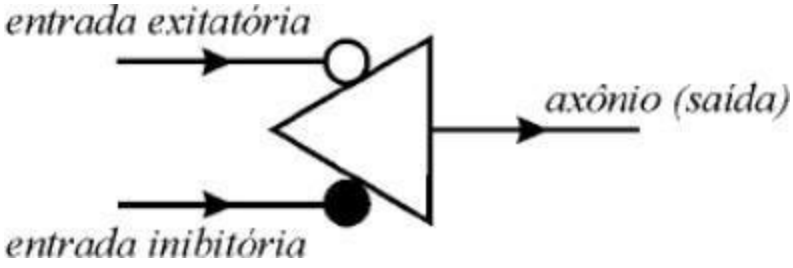

Figura 3: Modelo de RNA (ZAMBIASI, 2002).

O neurônio artificial, assim como o biológico possui uma ou mais entradas de sinais, e apenas um sinal de saída. A figura 4 mostra o modelo de McCulloch e Pitts da Equação do neurônio Artificial, onde temos as entradas X_i, e os pesos W_i, que são combinados usando uma função de soma ou limiar lógico, onde uma vez satisfeito o limiar da função, será produzida a saída do neurônio que seguirá para a cama seguinte se essa existir (BARRETO, 2002).

O resultado da saída de um neurônio é produto escalar das entradas pelos pesos:

$$\sum_{i=0}^{n} x_i w_i \qquad (4)$$

Após esta operação, os sinais de entradas passam a ser conhecidos como entradas ponderadas (ZAMBIASI, 2002).

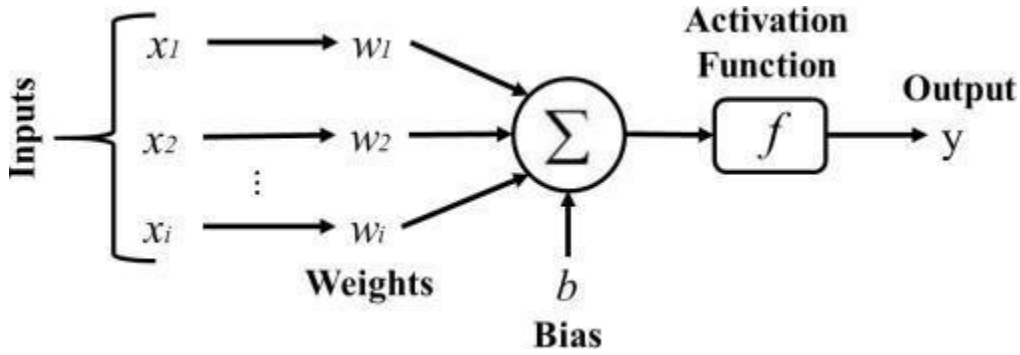

Figura 4: Modelo de RNA,(FONSECA, 2019)

2.2.3.1 A Arquitetura das Redes Neurais Artificiais

As RNA podem ter arquiteturas bem diferentes e únicas, podendo ter várias camadas e diferentes tipos de função de ativação. As RNAs que possuem apenas uma camada, como na figura 5, possuem apenas um nó entre a entrada e a saída, e são indicadas para problemas separáveis linearmente. Já as redes multicamadas, como na figura 6, possuem mais de uma camada entre as entradas e a saída, que são chamadas de camadas ocultas (ZAMBIASI, 2002; KASABOV, 1996).

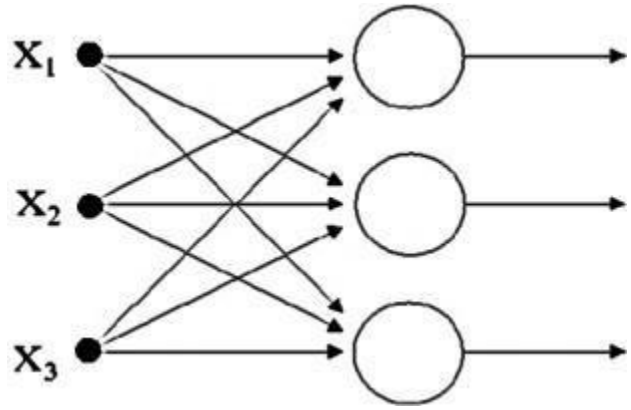

Figura 5: Arquitetura da RNA, (ZAMBIASI, 2002)

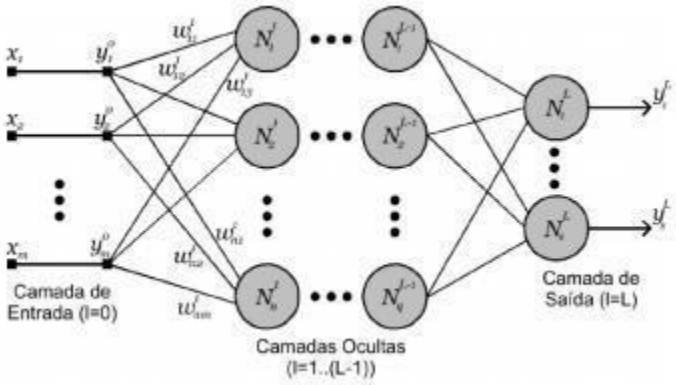

Figura 6: Arquitetura da RNA multicamadas (ZAMBIASI, 2002) O nós que compõem a RNA podem ter diferentes tipos de conexões:

- Feedforward: Onde a saída do neurônio na camada i não pode ser usada como entrada em nós das camadas anteriores (ZAMBIASI, 2002). Essa aplicação é típica para desenvolvimento de modelos não lineares que são utilizados para reconhecimento e classificação de padrões. Essa arquitetura é mostrada na figura 7.
- Feedback: Nessa configuração a saída do neurônio na camada i pode ser usada como entrada em camadas de índices anteriores. Essa configuração de rede associa um padrão de entrada com ele próprio, e são aplicadas geralmente em recuperação ou regeneração de um padrão de entrada, como no caso das Redes Neurais Recorrentes (ZAMBIASI, 2002). Esse modelo é mostrado na figura 8.

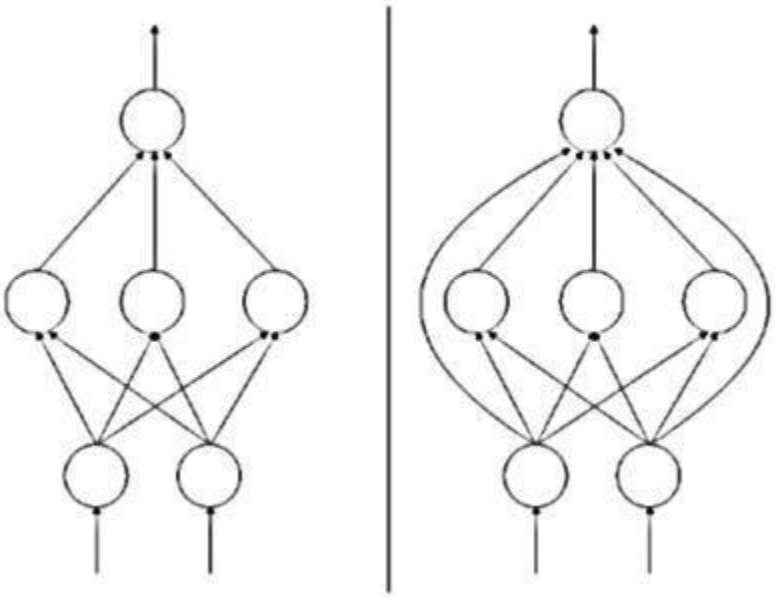

Figura 7: Arquitetura Feedforward (ZAMBIASI, 2002.)

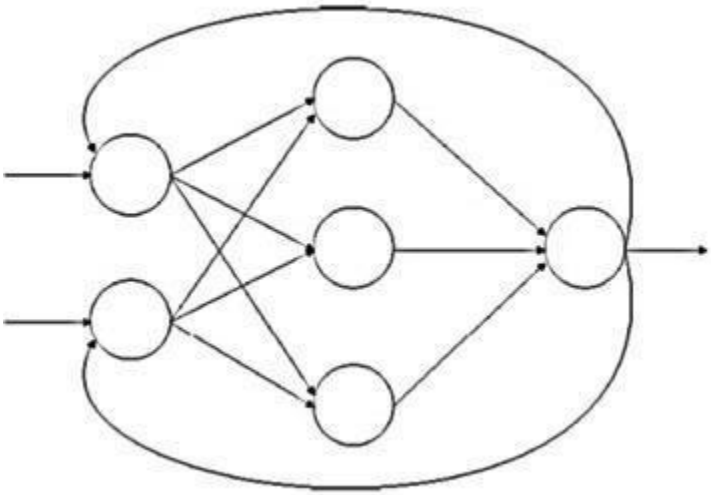

Figura 8: Arquitetura Feedback (ZAMBIASI, 2002.)

2.2.3.2 Algoritmo Backpropagation

É um algoritmo de treinamento baseado na correção de erros que compara a saída dada pela rede com a saída esperada no conjunto de treinamento. A partir dessa comparação, é calculado o erro, que é retornado camada por camada, ajustando os pesos e o bias para a interação seguinte, como mostrado na figura 9.

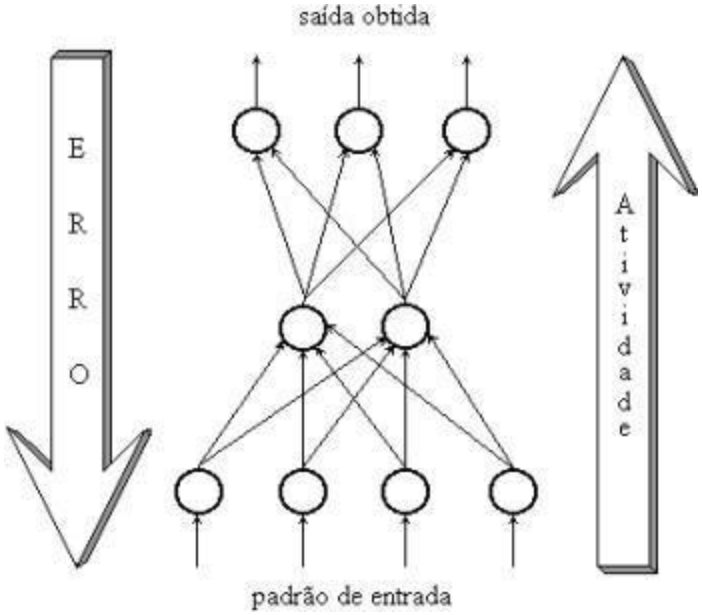

Figura 9 : Propagação de erro em uma RNA (ZAMBIASI, 2002.)

Dessa forma, o erro em cada neurônio j da camada de saída na interação t pode ser definido como a diferença entre os valores calculados e esperados,

$$e_j(t) = d_j(t) - y_j(t)$$

E pode-se obter a soma de erros quadrados de acordo com a eq.

$$\varepsilon(t) = \frac{1}{2}\sum_{j\epsilon s} e_j^2(t) \qquad (5)$$

Onde o S é o conjunto que agrega todos os N neurônios da camada de saída.

Alguns pontos devem ser chamados atenção para que o algoritmo tenha uma solução adequada. Os ajustes dos parâmetros e o valor da taxa de aprendizagem são os mais importantes, pois tem relação direta com o tempo necessário para a convergência da rede. Os critérios de parada também são de grande importância para que se evite esforços além do suficiente (HAYEK; NETWORK, 2004).

2.2.4 Avaliação dos Métodos de Classificação

Após o processo de classificação, os testes podem ser divididos em Falso Positivo (FP), que é o número de indivíduos não portadores da DRC classificados como portadores; Verdadeiro Positivo (VP), que é o número de portadores da DRC classificados de maneira correta como portadores da DRC; Falso Negativo (FN), que é quando o número de não

portadores da DRC é classificado como portadores e Verdadeiro Negativo (VN), que é quando não portadores são classificados como não portadores.

Essas métricas são utilizadas para quantificar o desempenho da metodologia utilizada, para avaliar o quão eficiente é o modelo e se os objetivos foram alcançados. Dessa forma, as medidas de desempenho utilizadas foram: especificidade (PROVOST; DOMINGOS,2000), Sensibilidade (PROVOST; DOMINGOS,2000), acurácia (SACHS, 2017) e a curva ROC (WANG; CHANG, 2010).

A sensibilidade é definida como a probabilidade de o teste fornecer resultado positivo, desde que o indivíduo seja portador da característica em questão. Dessa forma a sensibilidade define o número de indivíduos classificados corretamente (PROVOST; DOMINGOS, 2000). O cálculo da sensibilidade (S) é mostrado pela equação 2.2.35,

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

A especificidade, é a probabilidade do fornecimento de um resultado negativo pelo teste. Ou seja, o teste mede a proporção de pessoas sem determinada característica, indicando o poder do teste em identificar indivíduos não portadores da doença (PROVOST; DOMINGOS, 2000). O cálculo da especificidade (E) é dado pela equação 2.2.36.

$$E = \frac{VN}{VN + FP}$$

A acurácia (A), calcula o total de acertos baseado em todas métricas calculadas, calculando a partir dos indivíduos classificados corretamente e incorretamente (WANG; CHANG, 2010). O cálculo da acurácia é dado pela equação 2.2.37.

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

A partir da sensibilidade e da especificidade é possível calcular a curva ROC para todas as observações da amostra. O melhor ponto da curva ROC é no seu ponto superior esquerdo, onde a mesma adquire valor próximo de 1, o que é capaz de descrever o desempenho de um classificador à medida que o limite de discriminação varia (SACHS, 2017). A área sob a curva ROC (AUROC) é uma das medidas mais comuns usadas para avaliar o poder discriminativo do classificador (WANG; CHANG, 2010). A área representa a

probabilidade de classificação correta de uma determinada categoria.

2.3 Aprendizado Não-supervisionado

Neste tipo de aprendizado a máquina não exerce a função de supervisor para influenciar no processo de aprendizagem, o aprender por meio de investigação nos dados, a fim de obter conhecimentos sobre o mesmo. Dessa forma, possui a característica de observar e explorar a base de dados e como não há agente externo, não existe rótulo (JUSTO, 2016).

O processo consiste em agrupar uma quantidade de observações de acordo com alguma medida de dissimilaridade, em que observações que pertencem a um mesmo grupo sejam mais similares entre si mesmas e similares às observações que constituem os demais grupos. Os métodos mais comumente utilizados são o K-Means Clustering/K-Medoid e o Hierarchical Cluster (METZ, 2006).

2.3.1.1.1 Pré-processamento

As variáveis utilizadas possuem escalas diferentes, ou seja, os atributos que as caracterizam podem possuir valores de escala diferentes, o que pode influenciar no processo de classificação. Por esses motivos, muitas vezes é necessário realizar a normalização dos dados, a fim de manter os atributos dentro de uma faixa de valor específica para evitar que seja atribuída maior importância a atributos de escala maiores. (METZ, 2006).

O processo de normalização é feito antes do treino e teste dos algoritmos, Após a realização da normalização a média dos novos dados irá se concentrar em 0 e seu desvio padrão em 1. A equação abaixo define o processo de normalização, onde X representa o conjunto de observações, μ_x a média dos valores de X e σ_x o desvio padrão (GEMAN; BINNESTOCK; DOURSAT, 1992).

$$x_{novo} = \frac{X - \mu_x}{\sigma_x} \qquad (6)$$

2.4 Doença renal Crônica

A DRC é caracterizada pela redução progressiva e irreversível da função renal,

gerando um desequilíbrio metabólico e hidroeletrolítico do organismo e constituindo um impacto crescente na saúde pública, com elevação da morbimortalidade mundial (MARINHO et al., 2017; PROVENZANO, 2020).

De acordo com dados obtidos pelo Global Burden of Kidney Disease, a incidência e prevalência da doença renal a nível global aumentou em 88,76% e 86,95%, respectivamente, entre os anos de 1990 a 2016, e a mortalidade atribuída à DRC teve uma elevação de 41,5% entre os anos de 1990 e 2017 (XIE, 2018). No Brasil,a Sociedade Brasileira de Nefrologia (SBN) estimou um aumento de 54,1% de novos pacientes em diálise em 2018 em relação à 2009, totalizando 42.546 pacientes. Além disso, verifica-se que a prevalência de DRC em estágios iniciais ainda apresenta escassez de dados na literatura (NEVES, 2020).

Recentemente, um estudo de Alcalde & Kirsztajn (2018) demonstrou que as doenças renais e suas comorbidades representam aproximadamente 7,61% das internações e 12,97% dos gastos com internações no Brasil. Em 2015, os gastos com internações ocasionadas por diversas causas totalizaram R$ 13,8 bilhões e mais de R$ 2 bilhões com transplante renal e diálise. Nesse contexto, os dados apresentados representam um importante percentual de gastos nacionais em saúde e estima-se que o aumento permaneça constante independente das mudanças no perfil de desenvolvimento do país e do envelhecimento da população.

A TFG é amplamente utilizada na prática clínica para aferir a função excretora renal e o diagnóstico da DRC é baseado na detecção de uma TFG < 80mL/min/1,73m^2 por pelo menos três meses consecutivos. Já nos casos em que a TFG>80mL/min/1,73m^2 o diagnóstico pode ser realizado através da detecção de pelo menos um marcador de lesão renal,como albuminúria ou rins policísticos (BRASIL, 2014).

3 MÉTODOS

3.1 Amostra

Trata-se de um estudo transversal com 291 indivíduos atendidos em um Centro de Referência para tratamento de Doenças Renais na cidade de São Luís (MA), maiores de 18 anos, de ambos os sexos. A classificação da DRC foi realizada com base nas Diretrizes do Ministério da Saúde.

Os participantes foram esclarecidos quanto à finalidade do estudo e, quando de acordo, assinaram o Termo de Consentimento Livre e Esclarecido. O estudo foi aprovado pelo Comitê de Ética em Pesquisa com seres Humanos da Universidade Federal do Maranhão – UFMA

segundo parecer 2.035.753.

3.2 Critérios de inclusão

Foram incluídos indivíduos com idade igual ou superior a 18 anos, que estão em acompanhamento no Centro de Prevenção de Doenças Renais- HUUFMA.

3.3 Critérios de Exclusão

Foram excluídas pacientes gestantes ou com alguma incapacidade física que impossibilitasse ou comprometesse a coleta de dados. Pacientes com coletas de dados incompletos e os que não concordaram com o Termo de Consentimento Livre e Esclarecido.

3.4 Cálculo Amostral

A população estudada teve por base um universo finito representado por pacientes atendidos em centro de referência, entre os anos 2019 e 2021, totalizando 500 pacientes, segundo dados fornecidos pela instituição. Para a determinação do tamanho da amostra foi utilizado o cálculo proposto por Rea e Parker (2000), considerando um intervalo de confiança de 95%, um erro absoluto amostral de 5%, prevalência da DRC em adultos de 8,9%, (BARRETO, 2015) e adição de 10% para possíveis perdas ou recusas, conforme a equação a seguir:

$$n = \frac{Z\,a2.P.Q.N}{Z\,a2.P.Q + (N-1).E2} \qquad (7)$$

Onde: n = Amostra calculada (110 indivíduos); Zα – Variável correspondente ao nível de confiança; P – prevalência associada à causa; Q - variável associada aos indivíduos que não pertencem à categoria de interesse; N – Tamanho da população de referência; E - erro amostral.

3.5 Coleta de Dados

As medições foram realizadas por um único pesquisador com um mesmo instrumento calibrado. As medidas foram efetuadas em duplicata e as médias foram consideradas para a análise dos dados.

O peso foi aferido com uma balança eletrônica calibrada (Omron ® HBF 214 LA, Japão) com resolução de 0,1kg. A altura foi determinada com o auxílio de um estadiômetro transportável vertical com resolução de 0,1 cm (Sanny ®, Brasil). O Índice de Massa Corporal (IMC) foi obtido através da razão entre peso (Kg) e o quadrado da altura (m).

As circunferências foram medidas com uma trena antropométrica inelástica com precisão de 0,1cm (Seca ® 213, Hamburg, Germany). A circunferência da cintura (CC) foi aferida no ponto médio entre a última costela e a crista ilíaca na respiração mínima. Circunferência do braço direito e da panturrilha foram mensurados conforme descrito por Salmaso et al (2014). A circunferência do quadril seguiu os procedimentos descritos por Lohman (1986). A relação cintura estatura foi calculada através da razão entre a circunferência da cintura (cm) e a altura (cm) (ASHWELL; HSIEH, 2005)

Todas as análises bioquímicas foram realizadas no aparelho automatizado cobas 6000 (Roche) seguindo a metodologia descrita pelo fabricante. As coletas foram realizadas à vácuo com sistema de múltiplas coletas.

A pressão arterial sistólica (PAS) e a pressão arterial diastólica (PAD) serão aferidas com auxílio de aparelho monitor de pressão arterial de braço (OMRON®, modelo HEM 7130). A aferição e o ponto de corte utilizado seguiram as recomendações da VII Diretriz Brasileira de Hipertensão Arterial (SBC, 2017), na qual são considerados normais valores de PAS $\leq$ 120 mmHg e de PAD $\leq$ 80mmHg e alterados os valores de PAS > 120 mmHg e de PAD > 80 mmHg.

Tabela 1 – Tabela de variáveis utilizadas

Variáveis	Abreviação	Unidade
Sexo	-	-
Peso	-	Kg
Circunferência da Cintura	CC	cm
Relação Cintura Estatura	RCE	cm
Relação Cintura Quadril	RCQ	cm
Pressão Arterial Diastólica	PAD	mmHg
Pressão Arterial Sistólica	PAS	mmHg
Taxa de Filtração Glomerular	TFG	mL/min/1,73²

As variáveis utilizadas foram atribuídas no estudo com base em indicadores para avaliação nutricional e de saúde descritos na literatura e preconizados pela OMS. Os indicadores antropométricos são índices consolidados na avaliação de adiposidade central e

risco cardiovascular (BRASIL, 2014; JÉSUS, 2017; MANCINI, 2016; SALMASO, 2014).

3.6 Análise Estatística

A análise estatística foi feita através do software SPSS (Statistical Package for the Social Sciences, Inv., Chicago, IL, USA) versão 25. Foi feita a análise descritiva dos dados, bem como o teste de normalidade Kolmogorov-Smirnov. Para comparação entre os grupos, foram utilizados os testes t de Student e o teste Mann-Whitney U. Os resultados serão considerados estatisticamente significativos para $p < 0.05$.

3.7 Métodos de classificação

Para a criação do modelo de classificação foram utilizados a Regressão Logística e a Rede Neural Artificial. A arquitetura da RNA é composta por uma camada oculta e 7 variáveis de entrada. Nos treinamentos da rede, utilizou-se o algoritmo de retropropagação (backpropagation) com base na regra do momento. Os valores escolhidos para a taxa de aprendizagem foi de 0.3.

3.8 Divisão da base de treinamento e teste

O modelo possui 7 variáveis de entrada, descritas na tabela 1, e o resultado do modelo será a classificação da DRC em determinado paciente. Para evitar problemas de overfitting na RNA e na Regressão, foi utilizado o método k-fold-cross-validation (JUNG, 2018) como métrica de avaliação, em que os dados foram divididos em conjuntos de treinamento e teste, onde os dados são divididos igualmente em k segmentos iguais ou quase iguais. O treinamento e o teste são realizados por meio de k iterações (YADAV; SHUKLA, 2016). A base foi dividida aleatoriamente em 5 subconjuntos ($k = 5$), e o conjunto de treinamento e de teste tiveram uma proporção de 80% e 20%, respectivamente.

3.9 Validação

Após o processo de treinamento e teste, é preciso realizar a validação dos resultados encontrados. Para a análise de desempenho dos métodos propostos, foi utilizado as métricas de sensibilidade, especificidade, acurácia e a área sob a curva ROC, que são descritos na secção
2.2.4. Essas métricas têm como objetivo medir o desempenho do modelo proposto como

satisfatório ou não, além de auxiliar na identificação de pontos positivos e negativos para melhoria futura deste trabalho

4 RESULTADOS E DISCUSSÃO

A tabela abaixo apresenta a caracterização da amostra, separada através da classificação da DRC. Apenas a variável Relação Cintura Estatura não apresentou diferença estatisticamente significativa (p-valor < 0.005). A base de dados foi formada por 291 pacientes maiores de 18 anos e de ambos os sexos. Quanto ao sexo 75,5 % (n = 219) do sexo feminino e 24,5% (n = 71) do sexo masculino.

Tabela 2 Características antropométricas estratificada pela classificação da DRC

Variáveis	Não Portadores da DRC	Portadores da DRC	P - valor
Sexo			
Feminino	124	95	0.025
Masculino	26	45	
Peso	63,43 (+- 10,35	75.35(19.67)	0.001
Circunferência da Cintura	80.15 (9.86)	91.96(11.46)	0.001
Relação Cintura Estatura	0.513 (0.07)	0.569 (0.092)	0.061
Relação Cintura Quadril	0.823 (0.80)	0.88 (0.095)	0.009
Pressão Arterial Diastólica	72.01 (7.8)	84.97 (8.70)	0.000
Pressão Arterial Sistólica	111.59 (11.36)	135.707 (21.45)	0.001

Os modelos RNA e regressão foram aplicados para classificar pacientes portadores da DRC. A aplicação de ambos é composta por duas fases: treinamento e teste. No caso da regressão, no treinamento, as constantes de cada variável independente são calculadas, a fim de criar uma equação linear que classifica a DRC. Já para o modelo de rede neural, na fase de treinamento os parâmetros são ajustados até a função de custo alcançar o valor mínimo. A topologia da RNA é apresentada na tabela 3.

Tabela 3: Topologia da RNA

Topologia	Característica
Entradas	7
Camadas Ocultas	1
Neurônio na camada oculta	5
Taxa de aprendizagem	0.3
Função de ativação	ReLu
Tipo de RNA	FeedFooward
Algoritmo	Backpropagation

Os modelos tem como objetivo principal prever e classificar indivíduos portadores da DRC através de variáveis não invasivas. As variáveis de entrada foram descritas na seção 3.5. Na fase de testes da regressão e da RNA foram separadas 30% das amostras para serem utilizadas como teste. A escolha das amostras para treino e teste foi feita utilizando validação cruzada (k – folds cross validation). O desempenho dos modelos está descrito na tabela 4. Em relação às medidas de desempenho, a RNA foi superior à Regressão, com intervalo de confiança de 0.5, apesar de a área sob a curva ROC indiciar um desempenho semelhante entre os modelos.

Tabela 4: Resultados dos modelos

Parâmetros de classificação	AUROC (IC95%)	Acurácia	Sensibilidade	Precisão
RNA	0.94	87.7%	87%	87%
Regressão Logística	0.94%	85%	85%	85%

Conforme demonstrado na tabela 4, a RNA apresentou melhor desempenho em comparação com a regressão, embora a curva ROC indique uma semelhança na capacidade de classificação, como vemos na figura 11.

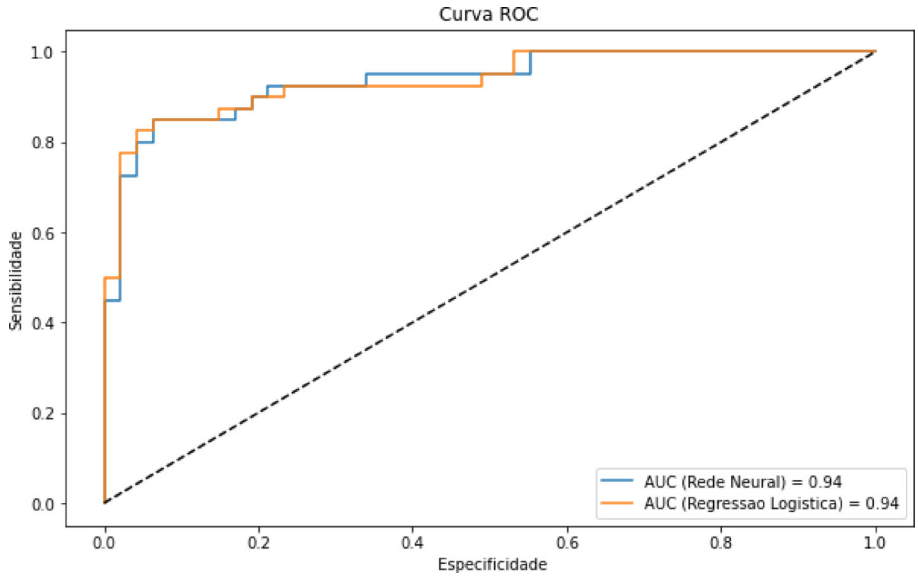

Figura 11: Curva ROC

Tanto a RNA quanto a regressão apresentaram alta sensibilidade para classificar pacientes portadores de DRC, a análise da curva ROC demonstrou que ambos os classificadores possuem um bom desempenho, pois se aproximou-se de 1 nesse parâmetro.

A aferição de medidas antropométricas têm sido empregada com diversas finalidades,

em especial, para identificação de risco, intervenção ou avaliação do impacto na saúde. Atualmente, tem se observado o desenvolvimento de importantes estudos envolvendo os indicadores antropométricos de saúde, nos quais métodos de estimação são propostos e pontos de cortes são definidos para populações diferentes (PIQUERAS et al. 2021).

A Circunferência da Cintura está associada ao acúmulo de gordura abdominal, que possui relação com um maior risco de mortalidade decorrente de complicações cardiovasculares (FONTES, 2018). As variáveis IMC, CC e RCE foram relacionadas como faotores de risco para a diminuição da TFG, além de estarem associadas ao aumento da adiposidade central, dessa forma possuem capacidade de auxiliar na identificação do risco cardiovascular associado à DRC (ODAGIRI, 2014).

Estudos evidenciam que a obesidade é uma das principais causas de comorbidades como resistência à insulina, hipertensão, hiperlipidemia, aterosclerose e intolerância à glicose, e destacam que a resistência insulínica pode estar envolvida na fisiopatologia da dislipidemia na doença renal crônica (MEYRIER, 2015).

A hipertensão arterial é um dos fatores de risco para a progressão da DRC, e possui relação com fatores como idade, maior rigidez cardiovascular e atividade do sistema nervoso simpático. Além disso, possui forte associação à diminuição da função renal e outras comorbidades como a obesidade e diabetes (LI et al, 2020; JEFFERS; ZHOU, 2017).

Níveis elevados de pressão arterial, através de medidas antropométricas, são associados diretamente ao excesso de peso corporal. Um estudo de Tran et al (2018), demonstrou que a Relação Cintura-Quadril (RCQ), quando maior que 0,95 em homens e 0,85 em mulheres, apresenta-se em indivíduos com maior predisposição às doenças cardiovasculares, como a hipertensão arterial sistêmica e suas complicações.

As variáveis Peso, CC, RCE e CQ foram identificadas com capacidade de discriminar os estágios da DRC e dessa forma auxiliar no monitoramento dessa patologia (SILVA, 2021). Dessa forma, a alta capacidade de classificação demonstrada pelos modelos aqui desenvolvidos tem estreita relação com as variáveis utilizadas como entradas.

Dessa forma, um dos fatores que pode ter influenciado no bom desempenho dos métodos aqui propostos foi a utilização de variáveis clínicas que já eram utilizadas na análise do estado de saúde dos pacientes (ROCHA, et al.,2019). Os modelos propostos apresentam vantagens bem visíveis, como o fato de poder ser aplicado em situações com limitações de recursos.

Os modelos de Regressão e RNA têm sido muito utilizados para tratar problemas

complexos do mundo real, e no caso da RNA, sua principal aplicação vem para lidar com modelos não lineares. A partir dos resultados encontrados, podemos ver que o modelo baseado em RNA teve um desempenho superior ao da regressão, porém o segundo tem um potencial maior e mais simples para aplicação no mundo real, devido ao fato da complexidade da estrutura da RNA.

Dessa forma, temos que nossos modelos obtiveram um bom desempenho para realizar a classificação dos doentes renais crônicos, apresentando-se então como uma alternativa de baixo custo para o acompanhamento e rastreamento da doença na população em geral.

5 CONCLUSÃO

A aplicação de técnicas de Regressão logística e RNA demonstraram serem capazes de realizar a classificação de pacientes portadores da DRC através de variáveis antropométricas as quais são consideradas parâmetros de baixo custo e fortemente associados aos fatores de risco da DRC.

Portanto, novos estudos direcionados com base nesses modelos são importantes para dar maior robustez aos modelos, para que estes venham a ser uma ferramenta utilizada no controle e diagnóstico da DRC em todos os níveis da assistência ao paciente com a implementação de medidas de prevenção e de tratamento da doença.

REFERÊNCIAS

ABREU, Mirhelen Mendes de et al. Health-related quality of life of patients recieving hemodialysis and peritoneal dialysis in São Paulo, Brazil: a longitudinal study. Value in health, v. 14, n. 5, p. S119-S121, 2011.

ALPAYDIN, Ethem. Introduction to machine learning. MIT press, 2020.

AMARAL, Fernando. Introdução à ciência de dados: mineração de dados e big data. Alta Books Editora, 2016.

BRASIL, Ministério da Saúde. DIRETRIZES CLÍNICAS PARA O CUIDADO AO PACIENTE COM DOENÇA RENAL CRÔNICA – DRC NO SISTEMA ÚNICO DE SAÚDE. Secretaria de Atenção à Saúde, v. 1, p. 1–37, 2014.

BRASIL. Diretrizes clínicas para o cuidado ao paciente com doença renal crônica-DRC no Sistema Único de Saúde. Secretaria de Atenção à Saúde, v. 1, p. 1–37, 2014.

BUSSAB, Wilton O.; MORETTIN, Pedro A. Estatística básica. 9. ed. São Paulo: Saraiva, 2017.

CARBONELL, Jaime Guillermo; MITCHELL, Tom Michael; MICHALSKI, Ryszard Stanislaw (Ed.). Machine learning: An artificial intelligence approach. M. Kaufmann., 1983.

FAVERO, L. P. et al. Análise de dados-modelagem multivariada para tomada de decisão 8ª ed. Rio de Janeiro: Ed. 2009.

FONTES, B. C. Efeitos da dieta hipoproteica sobre os perfis lipídico e antropométrico de pacientes com doença renal crônica em tratamento conservador. Brazilian Journal of Nephrology, v. 40, n. 3, p. 225–232, 2018.

HOSMER, David W.; JOVANOVIC, Borko; LEMESHOW, Stanley. Best subsets logistic regression. Biometrics, p. 1265-1270, 1989.

JAIN, Anil K.; DUBES, Richard C. Algorithms for clustering data. Prentice-Hall, Inc., 1988.
JAMES, Gareth et al. An introduction to statistical learning. New York: springer, 2013.

JEFFERS, Barrett W.; ZHOU, Duo. Relationship between visit-to-visit blood pressure variability (BPV) and kidney function in patients with hypertension. Kidney and Blood Pressure Research, v. 42, n. 4, p. 697-707, 2017.

JÉSUS, P. et al. Undernutrition and obesity among elderly people living in two cities of developing countries: Prevalence and associated factors in the EDAC study. clinical Nutr ESPEN, v. 21, p. 40–50, 2017.

JUNIOR, João Egidio Romão. Doença renal crônica: definição, epidemiologia e classificação. J. Bras. Nefrol., v. 26, n. 3 suppl. 1, p. 1-3, 2004.

JUSTO, Daniela Sbizera. Similaridade comportamental do consumo residencial

KAUFMAN, L.; ROUSSEEUW, P. J. Divisive analysis (program diana). Finding Groups in Data, p. 253-279, 2008.

KUNWAR, Veenita et al. Chronic Kidney Disease analysis using data mining classification techniques. In: 2016 6th International Conference-Cloud System and Big Data Engineering

(Confluence). IEEE, 2016. p. 300-305.

LI, Huihui et al. Visit-to-visit blood pressure variability and risk of chronic kidney disease: a systematic review and meta-analyses. PloS one, v. 15, n. 5, p. e0233233, 2020.

MANCINI, M. C. Diretrizes brasileiras de obesidade. Associação Brasileira Para O Estudo da Obesidade e da Síndrome Metabólica, v. 4, 2016. máquina. Rio de Janeiro: LTC, 2011.

MARGOTTO, P. R. Apostila: Entendendo BioestatÌstica Básica. Curso de Medicina da Escola Superior de Ciências da. Saúde/ESCS/SES/DF. 2012.

MARÔCO: João. Análise estatística com o PASW Statisitics. Report Number, Lda. Pêro Pinheiro: 2010.

MCCULLAGH, Peter; NELDER, John A. Generalized linear models. Chapman and Hall. London, UK, 1989.

METZ, Jean. Interpretação de clusters gerados por algoritmos de clustering hierárquico. 2006. Tese de Doutorado. Universidade de São Paulo.

MEYRIER, A. Nephrosclerosis: A Term in Quest of a Disease. Nephron, v. 129, n. 4, p. 276–282, 2015.

MITCHELL, Tom M. et al. Machine learning. 1997. Burr Ridge, IL: McGraw Hill, v. 45, n. 37, p. 870-877, 1997.

MONARD, Maria Carolina; BARANAUSKAS, José Augusto. Conceitos sobre aprendizado de máquina. Sistemas inteligentes-Fundamentos e aplicações, v. 1, n. 1, p. 32, 2003.

NEVES, P. D. M. DE M. Censo Brasileiro de Diálise: análise de dados da década 2009-2018. Braz. J. Nephrol, v. 42, n. 2, p. 191–200, 2020.

ODAGIRI, K. Waist to height ratio is an independent predictor for the incidence of chronic kidney disease. PloS one, v. 9, n. 2, p. e88873, 2014.

PAL, Debabrata; CHAKRABORTY, Chandan; MANDANA, K. M. Data mining approach for coronary artery disease screening. In: 2011 International Conference on Image Information Processing. IEEE, 2011. p. 1-6.

PERES, L. A. B. P.; BETTIN, T. E. Dislipidemia em pacientes com doença renal crônica. Rev. Soc. Bras. Clín. Méd, p. 10–13, 2015.

PINHO, N.A. et al. Prevalência e fatores associados à doença renal crônica em pacientes internados em um hospital universitário na cidade de São Paulo, SP, Brasil. Jornal Brasileiro de Nefrologia, v.37, n.1, p.91-97, 2015.

PIQUERAS, Paola Fiszman et al. Anthropometric indicators as a tool for diagnosis of obesity and other health risk factors: a literature review. Frontiers in Psychology, v. 12, p. 2618, 2021.

PROVENZANO, M. Contribution of Predictive and Prognostic Biomarkers to Clinical Research on Chronic Kidney Disease. International Journal of Molecular Sciences, v. 21, 2020.

PROVOST, Foster; DOMINGOS, Pedro. Well-trained PETs: Improving probability estimation trees. Raport instytutowy IS-00-04, Stern School of Business, New York University, 2000.

REMBOLD, S. M. Demographic profile of individuals with chronic renal disease from a multidisciplinary outpatient clinic of a university teaching hospital. Acta Paulista de

Enfermagem, v. 22, n. SPE1, p. 501–504, 2009.

SALMASO, F. V. Análise de idosos ambulatoriais quanto ao estado nutricional, sarcopenia, função renal e densidade óssea. Arquivos Brasileiros de Endocrinologia & Metabologia, v. 58, n. 3, p. 226–231, 2014.

SESSO, Ricardo Cintra et al. Brazilian chronic dialysis census 2014. Brazilian Journal of Nephrology, v. 38, n. 1, p. 54-61, 2016.

SILVA, S.B. et al. Cost comparison of kidney transplant versus dialysis in Brazil. Cadernos de saúde pública, v. 32, n. 6, 2016.

TABACHNICK, B. G.; FIDELL, L. S.; OSTERLIND, S. J. Using multivariate statistics. New York: Pearson, 2001.

TRAN, Nga Thi Thu et al. The importance of waist circumference and body mass index in cross-sectional relationships with risk of cardiovascular disease in Vietnam. PloS one, v. 13, n. 5, p. e0198202, 2018.

WANG, Zhanfeng; CHANG, Yuan-Chin Ivan. Marker selection via maximizing the partial area under the ROC curve of linear risk scores. Biostatistics, v. 12, n. 2, p. 369-385, 2011.

WITZ, K. Applied statistics for behavioral sciences. Journal of Educational Statistics, [S.l.], v. 15, n. 1, p. 84-87, 1990.

XIE, Y. Analysis of the Global Burden of Disease study highlights the global, regional, and national trends of chronic kidney disease epidemiology from 1990 to 2016. Kidney International, v. 94, n. 3, p. 567–581, 2018.

ZAMBIASI, S. P. Introdução às Redes Neurais Artificiais. Dissertação de Mestrado apresentada à Universidade Federal de Santa Catarina, Florianópolis – SC, 2002.

ZUBELLI, Fabrício Sander. Métodos de Inteligência Computacional para Clusterização de Consumidores no Setor de Energia Elétrica. 2017. Tese de Doutorado. Universidade Federal do Rio de Janeiro.

BRASIL, Ministério da Saúde. Portaria nº 1.675, de 07 de junho de 2018. Altera a portaria de Consolidação nº 3/GM/MS, de 28 de setembro de 2017, para dispor sobre os critérios para a organização, funcionamento e financiamento do cuidado da pessoa com Doença Renal Crônica - DRC no âmbito do Sistema Único de Saúde – SUS. Diário Oficial da República Federativa do Brasil. Brasília, DF, 07 de junho de 2018.

ERICKSON, Bradley J. et al. Machine learning for medical imaging. Radiographics, v. 37, n. 2, p. 505-515, 2017.

FERNANDES, Fernando Timoteo; FILHO, Alexandre Dias Porto Chiavegatto. Perspectivas do uso de mineração de dados e aprendizado de máquina em saúde e segurança no trabalho. Revista Brasileira de Saúde Ocupacional, v. 44, 2019.

WHO, World Health Organization. Noncommunicable diseases progress monitor 2020. Geneva, 2020.

SILVA, Ronaldo S. et al. Phenotypic Characterization of Chronic Kidney Patients Through Hierarchical Clustering. In: 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC). IEEE, 2021. p. 2451-2454.