Daniel Cavalcante de Oliveira

Uso de Redes Neurais Artificiais para predição de Síndrome Metabólica em Adolescentes

São Luís - MA

2022

Daniel Cavalcante de Oliveira

Uso de Redes Neurais Artificiais para predição de Síndrome Metabólica em Adolescentes

Monografia apresentada ao curso de Ciência da Computação da Universidade Federal do Maranhão como parte dos requisitos necessários para obtenção do grau de bacharel em Ciência da Computação.

Orientadora: Profa. Dra. Alcione Miranda dos Santos

São Luís - MA

2022

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

Cavalcante de Oliveira, Daniel.
Uso de Redes Neurais Artificiais para predição de Síndrome Metabólica em Adolescentes / Daniel Cavalcante de Oliveira. - 2022.
40 p.

Orientador(a): Alcione Miranda dos Santos.
Monografia (Graduação) - Curso de Ciência da Computação, Universidade Federal do Maranhão, São Luís - MA, 2022.

1. Adolescência. 2. Aprendizado de Máquina. 3. Dados desbalanceados. 4. Redes Neurais. 5. SM. I. Miranda dos Santos, Alcione. II. Título.

Daniel Cavalcante de Oliveira

Uso de Redes Neurais Artificiais para predição de Síndrome Metabólica em Adolescentes

Monografia apresentada ao curso de Ciência da Computação da Universidade Federal do Maranhão como parte dos requisitos necessários para obtenção do grau de bacharel em Ciência da Computação.

Aprovado em: 08/02/2022

Profa. Dra. Alcione Miranda dos Santos
Orientadora
Departamento de Saúde Pública - UFMA

Profa. Dra. Simara Vieira da Rocha
Examinadora
Departamento de Informática - UFMA

Profa. Msc. Debora Cristina Ferreira Lago
Examinadora
Departamento de Medicina III - UFMA

São Luís - MA
2022

Agradecimentos

Gostaria de agradecer primeiramente a Deus, por ter me dado o dom da vida, e por permitir que eu continue a respirar todos os dias, apesar de todas as tribulações e provações.

A minha família, em especial minha Mãe, minhas tias-avós, meu irmão e meu pai, que sempre me apoiaram em todos os meus sonhos e continuam me apoiando ate o dia de hoje.

A todos os professores que passaram pela minha vida até hoje, sem os ensinamentos adquiridos através de todos os níveis de ensino, jamais estaria onde eu estou hoje.

Ao Laboratório de Geotecnologias e Análise Espacial - GEOPRO, comandado pelo Prof. Dr. Maurício Eduardo Salgado Rangel, onde foi a minha segunda casa por mais de 5 anos dentro da UFMA, momentos muito especiais se passaram com os membros deste laboratório, em especial Lenilson Santiago por ter me convidado em 2016 a participar do laboratório, Suena (Susu) por toda a orientação ao longo de todos os dias e pela amizade mais que sincera, Tainan por ter sido mais que uma amiga, uma *friend*.

A todos os meus amigos da UFMA, tantas amizades se passaram por esses 6 anos de UFMA, que é até difícil mencionar todas, mas em especial, agradecimentos a Paulina, Ícaro, Maria, Emerson, Mônia, Asan, Stheffane, Arthur, Gabriela, Felícia, Dilssy, Luciano, Jemima, Leticia, Rodrigo, Lisle, Marília, dentre muitos outros, meu muito obrigado.

Ao grupo de Modelos Inteligentes do Programa de Pós Graduação em Saúde Coletiva, coordenado pela minha Orientadora, Profa. Dra. Alcione Miranda dos Santos, o qual tive a imensa sorte e prazer de achar através de uma lista de PIBIC, e ter sido aberto aos olhos da saúde coletiva e da maravilhosa estatística.

Ao CLC (Centro de Línguas e Cultura do Maranhão), na pessoa da Profa. Dra. Naiara Sales, o qual me deu minha oportunidade de entrar na docência, e descobrir que a docência, além da pesquisa, é o caminho que quero seguir para frente.

A todos que diretamente ou indiretamente, contribuíram para essa jornada, também os agradeço.

E a mim, que não desisti deste curso por mais de 6 anos, mas já não vejo a hora de ir para outros patamares!

Muito Obrigado!

"Think Different" (Steve Jobs)

Resumo

Síndrome Metabólica é um conjunto de fatores bioquímicos, fisiológicos, clínicos e metabólicos que se iniciam com a resistência periférica à insulina e que, quando concomitantemente presentes, elevam o risco de desenvolvimento de aterosclerose, eventos cardiovasculares, diabetes mellitus tipo 2 e mortalidade por qualquer causa. Em adolescentes, não há consenso sobre critérios diagnósticos e tratamento, por isso a importância de seu estabelecimento e padronização. Propõe-se determinar um modelo preditivo por meio das redes neurais artificiais para estimar a probabilidade de um adolescente possuir Síndrome Metabólica ou não. A base de dados usada no estudo foi a Coorte Retrospectiva RPS (Ribeirão Preto-SP, Pelotas-RS e São Luís-MA) usando dados somente da cidade de São Luís, composta por 2515 adolescentes entre 18 e 19 anos. Para definição de Síndrome Metabólica, usou-se o critério de Ferranti. Devido a baixa prevalência de Síndrome Metabólica na base, foi utilizado para o balanceamento dos dados, o algoritmo ROSE. As redes neurais artificiais do tipo *feed-forward* foram treinadas utilizando o algoritmo *back-propagation*. As métricas de avaliação das redes neurais artificiais utilizadas foram a acurácia, sensibilidade, especificidade, Kappa e a Curva ROC. O desempenho das RNA apresentadas neste trabalho mostraram bom desempenho para a predição de Síndrome Metabólica, em especial a RNA7 apenas com variáveis explicativas, apresentou acurácia de 81%, Kappa de 63%, Sensibilidade de 82% e Especificidade de 81%. Os resultados satisfatórios obtidos demonstram o potencial das redes neurais artificiais de predizer se um adolescente possuí Síndrome Metabólica ou não.

Palavras-chaves: Aprendizado de máquina. Dados desbalanceados. SM. Adolescência. Redes Neurais.

Abstract

Metabolic Syndrome is a set of biochemical, physiological, clinical and metabolic factors that start with peripheral insulin resistance and that, when concomitantly present, increase the risk of developing atherosclerosis, cardiovascular events, type 2 diabetes mellitus and mortality from any cause. In adolescents, there is no consensus on diagnostic criteria and treatment, so the importance of their establishment and standardization. We propose to determine a predictive model through artificial neural networks to estimate the probability of an adolescent to have Metabolic Syndrome or not. The database used in the study was the Retrospective RPS Cohort (Ribeirão Preto-SP, Pelotas-RS and São Luís-MA) using data only from the city of São Luís, consisting of 2515 adolescents between 18 and 19 years old. For definition of Metabolic Syndrome, we used the Ferranti criteria. Due to the low prevalence of Metabolic Syndrome in the database, the ROSE algorithm was used to balance the data. The feed-forward artificial neural networks were trained using the back-propagation algorithm. The metrics used to evaluate the artificial neural networks were accuracy, sensitivity, specificity, Kappa, and the ROC curve. The performance of the ANN presented in this work showed good performance for the prediction of Metabolic Syndrome, especially ANN7 with only explanatory variables, showed accuracy of 81%, Kappa of 63%, Sensitivity of 82% and Specificity of 81%. The satisfactory results obtained demonstrate the potential of artificial neural networks to predict whether an adolescent has Metabolic Syndrome or not.

Keywords: Machine Learning. Unbalanced data. Metabolic Syndrome. Adolescence. Neural Networks.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

SM - Síndrome Metabólica

IMC - Índice de Massa Corporal

HDL - Lipoproteína de alta densidade

LDL - Lipoproteína de baixa densidade

TG - Triglicerídios

RCE - Razão Cintura Estatura

CC - Circunferência da Cintura

PAS - Pressão Arterial Sistólica

PAD - Pressão Arterial Diastólica

Sumário

1 Introdução

A Síndrome Metabólica (SM) é definida como um conjunto de fatores bioquímicos, fisiológicos, clínicos e metabólicos que, quando concomitantemente presentes, elevam sobremaneira o risco de desenvolvimento de aterosclerose, eventos cardiovasculares, diabetes mellitus tipo 2 e mortalidade por qualquer causa. No Brasil, a Pesquisa Nacional de Saúde de 2013, iniciativa que levou em consideração dados de mais de 59.000 pessoas de todas as regiões do país, demonstrou prevalência de SM de 5,8% e 23,2% entre pessoas com idade entre 18-59 anos e com 60 anos ou mais, respectivamente (ALMEIDA et al., 2020).

A rápida urbanização, ingestão de excesso de energia, obesidade crescente e um predominante estilo de vida sedentário são os maiores responsáveis pelo aumento constante de síndrome metabólica e suas doenças associadas. A Síndrome metabólica é uma epidemia crescente e um grande fardo socio-econômico global. Ela foi acreditada como um biomarcador ativo em três grandes doenças não-comunicáveis: Doenças Cardiovasculares, Diabetes Mellitus Tipo II e câncer (KAKUDI; LOO; MOY, 2020).

Quando consideramos a síndrome metabólica na vida de adolescentes, percebemos que seus sintomas estão presentes, porém definir isso é incerto. Na comunidade acadêmica tal tema é polemizado e inconclusivo, devido a falta de uma definição universal dos pontos de corte para as variáveis da síndrome metabólica .Com inúmeros fatores de risco associados à SM em adolescentes, é de extrema importância identificar com máxima precisão a SM com a finalidade de intervir e minimizar alterações metabólicas futuras (ROSINI et al., 2015).

O intuito deste trabalho é unir os conhecimentos advindos da Ciência da Computação, como a utilização de inteligência artificial na forma das Redes Neurais Artificiais para criar uma rede neural capaz de predizer se adolescentes possuem ou não Síndrome Metabólica através dos sintomas e medidas comuns para aferição desta doença. Fazendo com que o diagnostico seja mais rápido e confiável à pacientes e doutores.

1.1 Trabalhos Relacionados

Poucos estudos têm demostrado o uso de técnicas de aprendizado de máquina para predizer a chance de um indivíduo desenvolver SM.

O estudo apresentado por Hirose et al. (2011), trás como objetivo predizer a incidência de síndrome metabólica em um período de seis anos utilizando Rede Neural Artificial e Análise de Regressão Logística Múltipla, baseado em fatores clínicos incluindo um índice de resistência a insulina calculado por um método de aferência homeostático. Utilizando dados da coorte de 410 professores homens japoneses da Universidade de Keio, entre 2000 e 2006, foram capazes de obter um resultado de 27% de sensibilidade para o modelo de regressão logística e 93% para a RNA, com especificidades de 95% e 91% respectivamente.

O estudo apresentado por Chen, Xiong e Ren (2014), tem o objetivo de avaliar o risco de síndrome metabólica utilizando modelos de inteligência artificial. A coorte utilizada foram 2074 pessoas entre 23 e 60 anos da Companhia Elétrica Lanzhou, na China, randomicamente selecionados. Foram utilizados dois modelos de IA: Regressão Logística de Componente Principal e Rede Neural de *Back-Propagation*. Foi utilizado também a curva ROC para avaliação dos resultados. O resultado para o modelo de Regressão Logística foi: 52% de sensibilidade, 92% de especificidade e Área debaixo da Curva de 88%. O modelo neural ofereceu um resultado substancialmente mais equilibrado, com sensibilidade de 88%, especificidade de 83% e Areá debaixo da curva de 90%, sendo este o modelo escolhido pelos autores como o melhor para classificação de risco de SM.

No estudo de Ivanović et al. (2016), o objetivo é simplificar a identificação da SM utilizando somente variáveis antropométricas (sexo, idade, IMC, RCE, Pressão Sistólica e Diastólica), o critério de IDF (ALBERTI et al., 2005) e redes neurais artificiais do tipo *Feed-forward*. Com uma coorte de 2928 indivíduos entre 18 e 76 anos da Sérvia, foi obtido um resultado de Valor de Predição Positivo de 85% e um Valor de Predição Negativo de 83%, sendo considerado pelos autores como uma solução adequada para predição de casos positivos e negativos de SM.

Todos esses estudos apresentaram seus resultados com populações consideradas adultas, a falta de aplicação das técnicas de aprendizado de máquina na população de adolescentes é uma deficiência que este trabalho vem suprir, visto que estas tem capacidade ótima de predizer chances de SM em adultos, independente do critério utilizado.

1.2 Objetivos

1.2.1 Objetivo geral

Determinar um modelo preditivo utilizando redes neurais artificiais para estimar a probabilidade de um adolescente desenvolver Síndrome Metabólica.

1.2.2 Objetivos específicos

- Determinar a prevalência de Síndrome Metabólica na amostra em estudo;
- Analisar a acurácia, sensibilidade e especificidade do modelo neural;
- Calcular o poder discriminatório do modelo neural.
- Avaliar a relevância dos indicadores antropométricos e cardiometabólicos na predição da Síndrome Metabólica (SM) em adolescentes.

1.3 Organização do trabalho

Este trabalho estará organizado em cinco seções distintas, compostas pela introdução, fundamentação teórica, a metodologia utilizada ao longo desta monografia, os resultados obtidos através dos experimentos e a conclusão com os pensamentos finais e trabalhos futuros.

2 Fundamentação teórica

Este capítulo apresenta a fundamentação teórica com os principais métodos e tecnologias utilizados bem como a contextualização da Síndrome Metabólica. Conceitos de aprendizado de máquina, aprendizado supervisionado, redes neurais artificiais, redes neurais *feed-forward* e seus componentes, dados desbalanceados e o algoritmo ROSE também serão fundamentados.

2.1 Síndrome Metabólica

A Síndrome Metabólica (SM) em população pediátrica é definido como o conjunto simultâneo da presença de três ou mais dos seguintes critérios: obesidade abdominal, dislipidemia (Triglicerídios elevado e/ou baixo HDL-C), resistência insulínica e/ou pressão sanguínea elevada (PIñA-AGUERO et al., 2018).

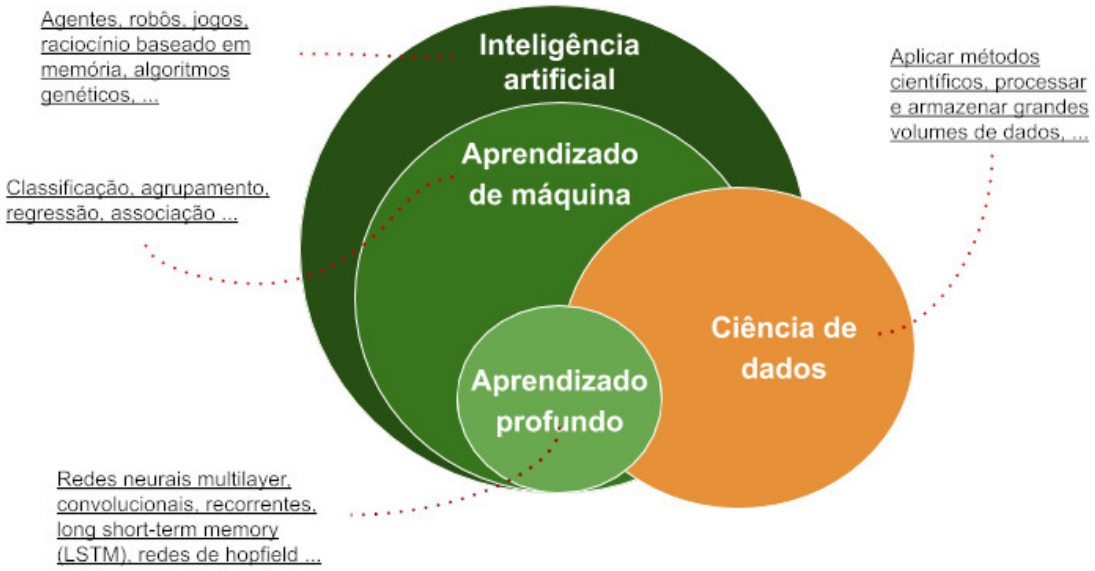
O aumento na prevalência de sobrepeso em adolescentes favorece a presença de SM, uma condição que agrega diferentes desordens metabólicas e que aumenta o risco de doença cardiovascular (DCV) e diabetes tipo 2 mellitus (DT2M) na fase adulta. Portanto, a identificação precoce dos fatores de risco e realizar tratamentos preventivos são estratégias desejáveis quando lidamos com uma epidemia de doenças cronicas não comunicáveis, que hoje são consideradas problemas de saúde pública (AGUDELO et al., 2014).

Para pacientes pediátricos que possuem Síndrome Metabólica não há um padrão e/ou definição aceita globalmente, apesar de mais de 40 tipos de critérios sugeridos foram publicados ao longo dos anos (PIñA-AGUERO et al., 2018). De acordo com Golley et al. (2006), existe uma prevalência de SM variando de 0 a 59% usando seis definições diferentes em uma mesma população de crianças pré-puberais com sobrepeso. Consequentemente, é extremamente difícil determinar qual critério seria mais apropriado para o ambiente clínico (PIñA-AGUERO et al., 2018). Além disso, faz-se necessário a identificação das variáveis que melhor discriminam a SM nesta população.

2.2 Aprendizado de Máquina

Aprendizado de máquina, também conhecida como Machine Learning (ML), é uma importante área da inteligência artificial, a qual permite a criação de algoritmos que ensinam determinada máquina a desempenhar tarefas. Na figura 1, podemos ver aplicações das diferentes áreas da Inteligência Artificial e Aprendizado de Máquina.

Figura 1 – Aplicações da Inteligência Artificial, Aprendizado de Máquina, Aprendizado Profundo e Ciência de Dados



Fonte: (DIAS, 2020)

As áreas de aplicações de ML são abundantes. Nas finanças, bancos utilizam dados passados para construir modelos a serem usados em aplicações de crédito, detecção de fraude e no mercado de bolsas. Na indústria, modelos de aprendizagem são usados para otimização, controle e controle de erros. Na medicina, programas de aprendizagem são utilizados em diagnósticos médicos. Nas telecomunicações, padrões de ligação são utilizados para otimização da rede e maximizar a qualidade do serviço. Na ciência, grandes quantidades de dados são utilizados na física, astronomia, e biologia podem ser analisados rapidamente com computadores (ALPAYDIN, 2020).

2.2.1 Aprendizado Supervisionado

Segundo (BONACCORSO, 2017), um cenário supervisionado é caracterizado pelo conceito de um professor ou supervisor, na qual sua tarefa principal é prover ao agente uma medida precisa do seu erro (diretamente comparável com os valores de saída).

Com algoritmos essa função é feita através de um *set* de treinamento feito com pares (entradas e saídas esperadas). A partir desta informação, o agente pode corrigir os parâmetros para reduzir a magnitude global da função de perda. Após cada iteração, se o algoritmo for flexível o suficiente e os dados estiverem coerentes, a acurácia geral irá crescer e a diferença entre o predito e o esperado ira estar perto de zero. Em um cenário supervisionado, o objetivo é treinar um sistema que deve trabalhar com amostras jamais vistas anteriormente, logo, se torna necessário permitir que o modelo desenvolva uma habilidade de generalização e evite um problema comum chamado *overfitting*, o qual causa

um problema de superaprendizado devido a um excesso de capacidade.

Aplicações de aprendizagem supervisionadas incluem:

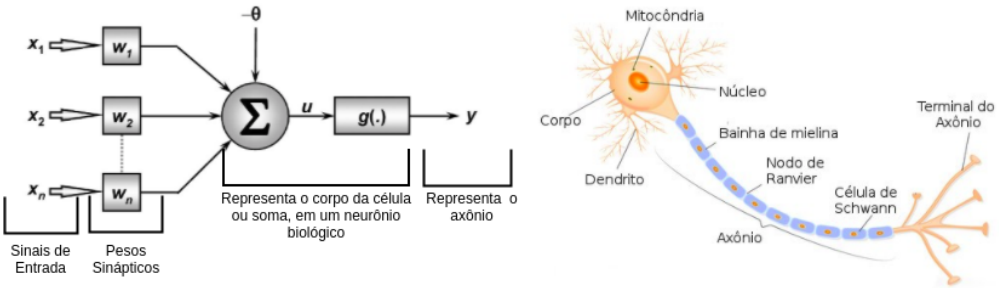
- Analise Preditiva baseada em regressão ou classificação categórica
- Detecção de SPAM
- Detecção de Padrões
- Processamento de Linguagem Natural
- Classificação automática de imagens
- Classificação de processos em sequencia (Musica ou Discursos)

Entre as técnicas do ML que utilizam aprendizado supervisionado, estão as Redes neurais artificiais que são abordadas na próxima seção.

2.3 Redes Neurais Artificiais

De acordo com Ciaburro e Venkateswaran (2017), a inspiração para as Redes Neurais foi a forma como o cérebro humano funciona. O cérebro humano pode processar gigantescas quantidades de informação usando dados mandados por sensos humanos. Tal processo é feito através de neurônios, cujo trabalho são passar sinais elétricos através deles e aplicando a logica *flip-flop*, como abrir e fechar portões para um sinal passar.

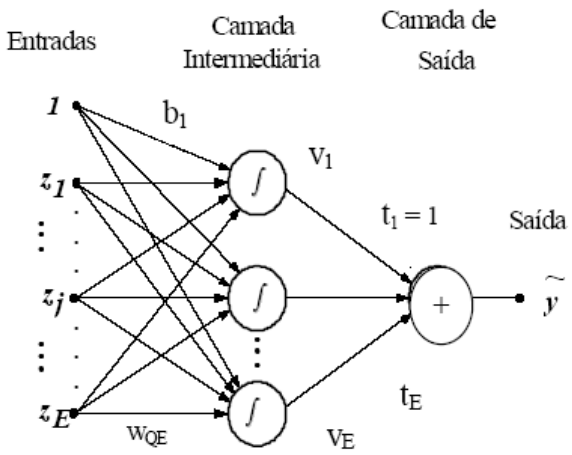
Figura 2 – Comparação entre Neurônio Biológico e Neurônio Artificial



Fonte: (VINICIUS, 2017)

Similar a estrutura neural biológica, a RNA define que o neurônio é a unidade central de processamento, o qual performa uma operação matemática para gerar uma saída de uma série de entradas. A saída de um neurônio é uma função da soma do peso das entradas mais o viés. Cada neurônio performa uma operação bem simples que envolve a ativação se o total de sinais recebidos excede o limite de ativação. Qualquer rede neural processando um *framework* tem a seguinte arquitetura:

Figura 3 – Exemplo de uma Rede Neural Artificial



Fonte: (HENDRIX, 2011)

Existe um conjunto de entradas, um processador, e um conjunto de saídas. As Entradas formam a *camada de entrada*, a *camada do meio* onde o processamento ocorre, é chamado de *camada oculta* e a (as) saída(as) formam a *camada de saída*.

2.3.1 Redes Neurais *Feed-forward*

As Redes Neurais *feed-forward* são modelos quintessenciais de aprendizagem profunda, segundo Goodfellow, Bengio e Courville (2016). O objetivo de uma rede neural *feed-forward* é aproximar algo a função f^*. Por exemplo, para um classificador, $y = f^*(x)$ mapeia uma entrada x a uma categoria y. A rede neural *feed-forward* define um mapeamento $y = f(x; \theta)$ e aprende o valor dos parâmetros θ que resultam na melhor função de aproximação.

Esses modelos são chamados *feed-forward* porque as informações fluem através da função sendo avaliada por x, por meio do intermédio computacional usado para definir f, e finalmente a saída y. Não existem conexões recorrentes nos quais as saídas dos modelos são alimentadas dentro de si.

Essas redes são denominadas **redes**, por que são tipicamente representadas por compor a junção de varias funções. O modelo é associado com um gráfico acíclico direto descrevendo como as funções são compostas juntas. Tendo como exemplo três funções diferentes $f^{(1)}$, $f^{(2)}$, e $f^{(3)}$ conectadas em uma corrente, para formar $f(x) = f^{(3)}(f^{(2)}(f^{(1)}(x)))$. Essas estruturas em corrente são comumente usadas como estruturas de redes neurais, Nesse caso $f^{(1)}$ é chamada de **primeira camada** da rede, $f^{(2)}$ é chamado de **segunda camada** e assim por diante. O tamanho total da corrente da a profundidade do modelo (GOODFELLOW; BENGIO; COURVILLE, 2016).

A camada final de uma rede neural *feed-forward* é chamada de **camada de saída**. Durante o treino da rede neural, nos direcionamos $f(x)$ a igualar $f^*(x)$. Os dados de treinamento nos trazem exemplos com ruídos aproximados de $f^*(x)$ avaliados em diferentes pontos de treinamento. Cada exemplo de x é acompanhado de um rótulo $y \approx f^*(x)$. Os exemplos de treino diretamente especificam o que a camada de saída deve fazer a cada ponto x: deve produzir um valor próximo de y. O comportamento de outras camadas não é diretamente especificado pelos dados de treinamento (GOODFELLOW; BENGIO; COURVILLE, 2016).

O algoritmo de aprendizagem deve decidir como usar essas camadas para produzir a saída desejada, mas os dados de treinamentos não dizem o que cada camada individual deve fazer. Ao invés, o algoritmo de treinamento deve decidir como usar essas camadas para melhor implementar uma aproximação de f^*. Por causa dos dados de treinamento não mostrarem a saída desejada de cada uma das camadas, estas são chamadas de camadas ocultas (GOODFELLOW; BENGIO; COURVILLE, 2016).

Um problema único as redes neurais *feed-forward* segundo Goodfellow, Bengio e Courville (2016), é a escolha do tipo de unidades ocultas a serem utilizadas nas camadas ocultas do modelo. O desenho das unidades é uma areá extremamente ativa de pesquisa e ainda ano existe muitos princípios teóricos definitivos.

Unidades lineares retificadas são uma excelente escolha padrão de unidade oculta. Outros tipos de unidades ocultas estão disponíveis. Isso pode dificultar quando se determina qual tipo utilizar. O processo de desenho da camada oculta é de tentativa e erro, inferindo qual tipo de cada oculta for utilizar pode funcionar bem e depois treinar a rede com esse tipo de cada oculta e avaliar sua performance nos dados de teste.

2.3.2 Desenvolvimento da RNA

2.3.2.1 Definição da arquitetura da RNA

Determinar a arquitetura de uma rede neural é um processo chave. A palavra arquitetura nessa instancia, se refere a estrutura geral de uma rede: quantas unidades deve ter e como essas unidades estarão conectadas uma a outra (GOODFELLOW; BENGIO; COURVILLE, 2016).

A grande maioria das redes neurais está organizada em grupos de unidades chamado de camadas. A maior parte das arquiteturas de redes neurais arranja essas camadas em uma estrutura em corrente, com cada camada sendo uma função da camada precedida. Nessa estrutura, a primeira camada é definida por

$$h^{(1)} = g^{(1)}(W^{(1)}x + b^{(1)}); \tag{2.1}$$

a segunda camada é definida por

$$h^{(2)} = g^{(2)}(W^{(2)}x + b^{(2)}); \tag{2.2}$$

e assim por diante.

Nas arquiteturas baseadas em correntes, a principal consideração arquitetural é escolher a profundidade da rede e o tamanho de cada camada. Uma rede com somente uma cada oculta é suficiente para encaixar os dados de treino. Redes mais profundas usualmente utilizam poucas unidades por camada e poucos parâmetros, assim como generalizam o seus dados de teste, mas tendem a ser mais difíceis de otimizar. Uma arquitetura ideal de rede para uma determinada tarefa deve ser encontrada através de experimentação e monitoramento do valor do erro nos dados de validação (GOODFELLOW; BENGIO; COURVILLE, 2016).

2.3.2.2 Funções de Ativação

Segundo Sharma, Sharma e Athaiya (2017), as funções de ativação são especialmente utilizadas nas redes neurais artificiais para transformar um sinal de entrada em um sinal de saída, no qual é alimentado como uma entrada para a próxima camada da operação. Em uma RNA, é calculado a soma dos produtos das entradas e seus correspondentes pesos para aplicar a função de ativação e conseguir a saída de uma determinada camada, suprindo como entrada para a próxima camada.

Caso uma função de ativação não seja utilizada em uma rede neural, logo o seu sinal de saída seria somente uma função simples linear. Sem as funções de ativação, a rede neural se torna um modelo de regressão linear com performance limitada. Por isso se torna necessário aplicar as funções de ativação dentro das Redes Neurais, para que a rede chegue ao seu potencial máximo sendo capaz de lidar com conjuntos de dados não-lineares, modelos com múltiplas camadas, arquiteturas complexas e afins (SHARMA; SHARMA; ATHAIYA, 2017).

A função sigmóidal, de acordo com Lederer (2021), é limitada e diferenciável que não diminuí e tem exatamente um ponto de inflexão. Uma função sigmóidal é uma função suave com uma curva em formato de "S". A ativação do Sigmoide tem um longa tradição na teoria e na prática das redes neurais. Uma motivação tem sido os padrões de ativação em forma de sigmoide observados em neurociência.

A função utilizada neste trabalho é a função logística sigmoidal que é representada por

$$z = \frac{1}{1 + e^{-z}} \tag{2.3}$$

A ativação logística é comumente utilizada para modelos preditivos que tem probabilidade como saída. Como a probabilidade de algo existir é definida entre 0 e 1, a ativação logística é a escolha correta devido ao seu alcance.

2.3.2.3 Algoritmo *Backpropagation*

Quando utilizamos uma rede neural *feed-forward* para aceitar uma entrada x e produzir uma saída $\hat{y}$, a informação flui pela rede. A entrada x provê a informação inicial que propaga para as unidades ocultas de cada cama e finalmente produz $\hat{y}$. Isto é chamado de propagação direta (ou *forward propagation*. Durante o treinamento, a propagação direta pode continuar até produzir um custo escalar $J(\theta)$. O algoritmo *back-propagation* (RUMELHART; HINTON; WILLIAMS, 1986), comumente chamado de *backprop*, permite que a informação sobre o custo flua de volta pela rede para computar o gradiente (GOODFELLOW; BENGIO; COURVILLE, 2016).

Computar uma expressão analítica para o gradiente é simples e direto, mas numericamente avaliar tal expressão pode ser computacionalmente caro. O algoritmo de *back-propagation* realiza isso usando um procedimento simples e eficiente.

As equações de *back-propagation* nos provêm com uma maneira de computar o gradiente da função de custo, em forma de algoritmo nós temos:

- **Entrada x:** Seta a ativação correspondente a^1 para a camada de entrada.
- ***Feed-Forward*:** Para cada $l = 2, 3 \ldots, L$ computar $z^l = w^l a^{l-1} + b^l$ e $a^l = \sigma(z^l)$
- **Erro de saída** δ^L : Computar o vetor $\delta^L = \nabla_a C \odot \sigma'^{(z^L)}$.
- ***Backpropagate the error*:** Para cada $l = L-1, L-2, \ldots, 2$ computar $\delta^l = ((w^{l+1})^T \delta^{l+1}) \odot \sigma'^{(z^l))}$
- **Saída:** O gradiente da função custo é dado por $\frac{\vartheta C}{\vartheta w^l_{jk}} = a^{l-1}_k \delta^l_j$ e $\frac{\vartheta C}{\vartheta b^l_j} = \delta^l_j$

Examinando o algoritmo acima podemos perceber por que é chamado de *back-propagation*. Os erros dos vetores δ^l são computados para trás, começando da camada final. Pode parecer peculiar a forma que se é navegado para trás na rede, mas se pensarmos no "porque", descobriremos que o movimento para trás é uma consequência do fato de que o custo é a função de saídas da rede. Para entender melhor como o custo varia com os pesos iniciais e viéses, precisamos repetidamente aplicar a regra da cadeira, trabalhando para trás através das camadas para a compreensão do funcionamento do algoritmo (NIELSEN, 2015).

2.3.2.4 Amostra de treino e de teste

Segundo Brownlee (2020), o procedimento de divisão dos conjuntos treino-teste, é usado para estimar a performance dos algoritmos de aprendizado de maquina quando eles são utilizados para fazer predições em dados não utilizados para treinar o modelo. Esse processo pode ser usado para qualquer problema de classificação ou regressão, também como para qualquer algoritmo de aprendizagem supervisionado.

O procedimento envolve pegar uma base de dados e dividir em dois sub-conjuntos. O primeiro sub-conjunto é usado para treinar o modelo, e é referido como conjunto de treino. O segundo sub-conjunto não é utilizado para treinar o modelo, o elemento de entrada da base de dados, é provido para o modelo, então predições são feitas e comparadas com os valores esperados. Este segundo sub-conjunto é chamado de conjunto de teste (BROWNLEE, 2020).

A idéia de "suficientemente grande"é específica para cada problema de modelagem preditiva. Isso significa que há dados suficientes para dividir o conjunto de dados em treino e conjuntos de dados de teste e cada um dos conjuntos de dados de treino e teste são representações adequadas do domínio do problema. Isto requer que o conjunto de dados original seja também uma representação adequada do domínio do problema. Uma representação adequada do domínio do problema significa que há registros suficientes para cobrir todos os casos comuns e os casos mais incomuns no domínio. Isto pode significar combinações de variáveis de entrada observadas na prática. Pode exigir milhares, centenas de milhares, ou milhões de exemplos (BROWNLEE, 2020).

Por outro lado, o procedimento de treino-teste não é apropriado quando o conjunto de dados disponíveis é pequeno. A razão é que quando o conjunto de dados é dividido em conjuntos de treinamento e testes, não haverá dados suficientes no conjunto de dados de treinamento para que o modelo aprenda um mapeamento eficaz de entradas para saídas. Também não haverá dados suficientes no conjunto de dados de teste para avaliar efetivamente o desempenho do modelo. O desempenho estimado pode ser excessivamente otimista (bom) ou excessivamente pessimista (ruim) (BROWNLEE, 2020).

O procedimento tem um parâmetro principal de configuração, que é o tamanho dos conjuntos de treino e teste. Isto é comumente expresso como uma porcentagem entre 0 e 1 para o treino ou para os conjuntos de dados de teste. Por exemplo, um conjunto de treinamento com o tamanho de 0,67 (67%) significa que a porcentagem restante 0,33 (33%) é atribuída ao conjunto de teste. Não há uma porcentagem de divisão ideal (BROWNLEE, 2020).

A porcentagem dividida escolhida tem que atender aos objetivos do projeto com considerações que incluam: Custo computacional no treinamento do modelo, custo computacional na avaliação do modelo, representatividade do conjunto de treinamento,

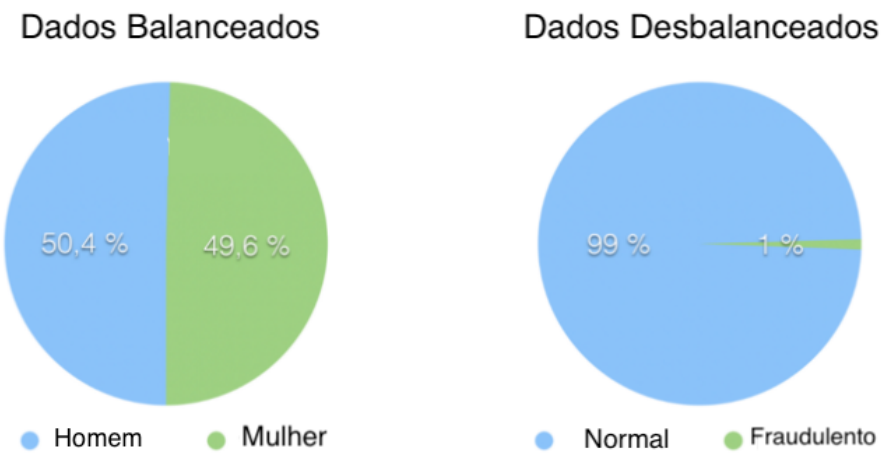
representatividade do conjunto de testes. No entanto, as porcentagens de divisão comuns incluem:

- Treino: 80%, Teste 20%
- Treino: 70%, Teste 30%
- Treino: 50%, Teste 50%

2.3.3 Dados desbalanceados

Dados desbalanceados são prevalentes em uma multitude de áreas e setores. O desafio aparece quando os algoritmos de aprendizagem de maquina tentam identificar esses casos raros em grandes bancos de dados. Devido a disparidade das classes nessas variáveis, o algoritmo tende a categorizar a partir da classe com mais instancias, a classe da maioria, enquanto ao mesmo tempo dando uma falsa impressão de modelo com alta acurácia. Tanto a inabilidade de prever raros eventos, a classe minoritária, e a falsa acurácia destoam dos modelos preditivos construídos (LAHERA, 2019).

Figura 4 – Exemplo de um Banco de Dados Desbalanceado



Fonte: Adaptado de Lahera (2019)

Na figura 4, podemos ver com clareza, que o gráfico mostrado a esquerda, representa um banco de dados balanceado, com proporções normais para não termos viés no algoritmo. Já o gráfico da direita, representa um banco de dados onde 99% de seus dados são de determinada classe, e 1% de outra classe.

Independente do algoritmo a ser escolhido, quando o modelo é treinado encima de um banco de dados desbalanceado, eles terão resultados ruins quando o modelo for colocado para testes com dados generalizados, além do modelo final estar com viés, de ter aprendido

mais exemplos somente de uma classe, tendo dificuldades de enxergar padrões. Para isso, existem soluções como *undersampling* e *oversampling* os quais diminuem os resultados da classe majoritária e aumentam os casos da classe minoritária, respectivamente. Porém para os propósitos deste trabalho, foi utilizado o algoritmo ROSE.

2.3.4 Algoritmo ROSE

ROSE (*Random Over-Sampling Examples* (LUNARDON; MENARDI; TORELLI, 2015) ajuda na tarefa da classificação binaria na presença de classes raras. Ele produz uma amostra de dados sintética, possivelmente balanceada simulada de acordo com uma abordagem *smoothed-bootstrap*.

Denotado por y como a resposta binaria e por x como um vetor de preditores numéricos observados em n indivíduos i, $(i = 1, \ldots, n)$, são gerados exemplos sintéticos com a classe denominada $k, (k = 0, 1)$, a partir de uma estimativa do kernel da densidade condicional $f(x|y = k)$. O kernel é uma função de produto Normal no centro de cada x_i com a matriz de covariância diagonal H_k. Neste caso, H_k é a matriz de suavização assimptóticamente ótima, assumindo que ela está sob o pressuposto da normalidade multivariada.

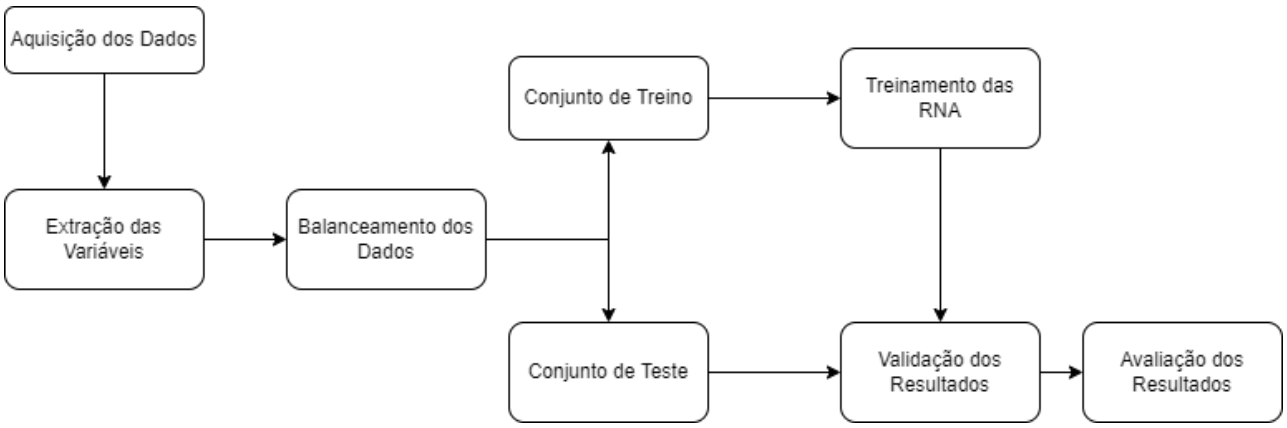
Essencialmente, ROSE seleciona uma observação pertencente a classe k e gera novos exemplos no seu *bairro*, onde a largura do *bairro* é determinada por H_k. O usuário é permitido a diminuir H_k variando os argumentos *h.mult.majo* e *h.mult.mino* dentro do algoritmo. O Balanceamento é regulado por argumentos, isto é, a probabilidade de gerar exemplos da classe $k = 1$.

Na sua forma atual, métodos baseados em kernel podem ser aplicados somente em dados contínuos. No entanto, ROSE inclui uma combinação de *over* e *under-sampling* como um caso especial quando H_k tende a zero, a presunção de continuidade pode ser evitada usando um kernel de distribuição degenerada para desenhar exemplos categóricos sintéticos. Basicamente, se $j-$th componente de x_i é categórico, um clone sintético de x_i terá $j-$th componentes de mesmo valor de $j-$th que é componente de x_i.

3 Aspectos Metodológicos

A metodologia proposta para este trabalho é composta por 6 etapas: aquisição da base de dados, definição das variáveis de entrada da RNA, balanceamento dos dados, divisão da amostra de treino e teste, treinamento e avaliação dos resultados da rede. Um resumo deste processo pode ser observado na Figura 5. As etapas citadas serão descritas nas seções a seguir.

Figura 5 – Metodologia Proposta



3.1 Base de dados

A base de dados utilizada neste estudo é proveniente da Coorte de Estudo retrospectivo RPS (Ribeirão Preto-SP, Pelotas-RS e São Luís-MA), cujo principal objetivo foi investigar os determinantes de saúde na adolescência, com foco nos desfechos referentes à nutrição e à composição corporal, bem como precursores de doenças crônicas complexas, saúde mental e capital humano. Para este estudo, serão considerados apenas os dados da terceira fase da Coorte RPS no município de São Luís-MA.

A primeira fase da Coorte, foi conduzida em dez hospitais da cidade, públicos e privados, de março de 1997 a fevereiro de 1998. Foram incluídos os nascimentos ocorrido em São Luís-MA neste período, correspondendo a 96,3% dos nascimentos. Os nascimentos não hospitalares e os ocorridos em hospitais com menos de 100 partos por ano foram excluídos.

Os nascimentos foram selecionados por amostragem sistemática com estratificação proporcional de acordo com o número de nascimentos em cada maternidade, de um em cada sete partos. Obteve-se um total de 2.831 nascimentos. Entretanto, os não residentes em São Luís, gemelares e natimortos foram excluídos, a amostra final foi de 2.443 nascimentos,

com 5,8% de perdas devido a recusas ou alta precoce (SIMõES et al., 2020).

Essa coorte foi acompanhada aos 7-9 anos (segunda fase) e novamente aos 18-19 anos (terceira fase). A amostra que compõe a terceira fase da coorte RPS é constituída por adolescentes de 18 a 19 anos, avaliados no ano de 2016, composta por 684 adolescentes. Com o objetivo de aumentar o poder das análises, e para prevenir perdas futuras, a coorte foi aberta para incluir outros indivíduos nascidos em São Luís, Maranhão no ano de 1997. Desta forma, a amostra final foi composta por de 2.515 adolescentes. A descrição dos métodos utilizados na coorte de nascimento está detalhada em Simões et al. (2020).

3.2 Variáveis em estudo

Na tabela 1, são apresentadas variáveis consideradas neste estudo, as quais foram obtidas a partir da base de dados da coorte RPS.

Tabela 1 – Aferição dos Dados das Variáveis - Coorte RPS

Variáveis	Método de Coleta
Idade (anos) e Sexo (M/F)	Por meio do Questionário Padrão da Coorte RPS
Pressão Arterial Sistólica e/ou Pressão Diastólica	Calculado com base na média das três medidas aferidas no aparelho Omron HEM 742INT (Omron, São Paulo, Brasil), obtidas após cinco minutos de repouso.
Peso (Kg)	Avaliado na balança de alta precisão acoplada ao equipamento BOD POD *Gold Standard* da marca COSMED (COSMED *Metabolic Company*, Roma, Itália).
Altura (cm)	Estadiômetro AlturaExata (Belo Horizonte, Brasil).
Circunferência (cm)	Pletismografia por deslocamento de ar no equipamento BOD POD *Gold Standard*
Glicemia (mg/Dl), TG (mg/Dl), HDL (mg/Dl)	Exames Bioquímicos.

Fonte: Adaptado de Simões et al. (2020) pelo autor.

A glicemia aleatória, colesterol total, *high-density lipoprotein cholesterol* (HDL-c) e triglicerídeos foram dosados pelo método colorimétrico enzimático automatizado pelo equipamento Cobas c501 da Roche.

O índice de massa corporal (IMC) foi obtido por meio da razão: peso corporal (kg)/altura (m^2). Para avaliar a proporção de gordura central pela estatura dos participantes

foi considerado a variável razão cintura-estatura (RCE), calculada dividindo-se a medida da circunferência da cintura (cm) pela altura (cm).

Apesar da variedade de critérios, para este estudo foram considerados os três critérios mais populares para definição de SM: IDF (ALBERTI et al., 2005), Cook (COOK et al., 2003) e Ferranti (FERRANTI et al., 2004). A partir desses critérios foi estimada a prevalência de SM na amostra em estudo (4,8% (IDF), 4,1% (Cook) e 10,2% (Ferranti)).

Em problemas de classificação, os conjuntos de treinamento e de teste devem ser representativos do problema que se pretende estudar. Quando a quantidade de dados disponíveis é pequena, cuidados devem ser tomados para que o conjunto de treinamento e de teste tenham representatividade do problema em estudo (SANTOS et al., 2005). Assim, optou-se para o critério de Ferranti, visto que este critério forneceu maior prevalência de SM. Na tabela 2 são apresentados os critérios de decisão de Síndrome Metabólica. Para considerar que o indivíduo tem síndrome metabólica, ele deverá apresentar três ou mais critérios da SM.

Tabela 2 – Critérios de Definição de SM de Ferranti et al. (2004)

Critérios	Definição Pediátrica Proposta
Circunferência da Cintura	‘> Percentil 75% para idade e gênero
Glicemia	>= 6.1 mmol/L
Pressão Sistólica	> Percentil 90% para idade, gênero e altura
Pressão Diastólica	> Percentil 90% para idade, gênero e altura
HDL	< 1.3 mmol/L (para meninos de 15 a 19 anos, <1.17 mmol/L
Triglicerídios	>= 1.1 mmol/L

Fonte: (FERRANTI et al., 2004)

3.3 Implementação da RNA

Conforme apresentado, o objetivo principal deste estudo é propor um modelo estatístico capaz de predizer a chance de um adolescente ter SM. Assim, definimos a variável resposta de interesse y da seguinte forma

$$y = \begin{cases} 1, & \text{se o adolescente possui SM} \\ 0, & \text{caso contrário} \end{cases}$$

Para construção do modelo preditivo, diferentes RNA *Feed-forward* foram consideradas neste estudo. As RNAs avaliadas consideraram além das componentes da SM, idade, sexo, IMC, peso, altura e RCE como variáveis de entrada e presença de SM (sim/não) como variável de saída da rede. Abaixo são apresentadas as redes implementadas no estudo utilizando a linguagem R e o programa RStudio.

- **RNA 1**: Glicemia, Circunferência da Cintura, HDL, Triglicerídios, Pressão Sistólica, Pressão Diastólica.
- **RNA 2**: Glicemia, Circunferência da Cintura, HDL, Triglicerídios, Pressão Sistólica, Pressão Diastólica, Idade e Sexo.
- **RNA 3**: Peso, Altura, Pressão Sistólica, Pressão Diastólica, Idade e Sexo.
- **RNA 4**: Idade, Sexo, Circunferência da Cintura, Pressão Sistólica, Pressão Diastólica.
- **RNA 5**: Idade, Sexo, IMC, Pressão Sistólica, Pressão Diastólica
- **RNA 6**: Idade, Sexo, RCE, Pressão Sistólica, Pressão Diastólica, Peso.
- **RNA 7**: Idade, Sexo, IMC, Circunferência da Cintura, Pressão Sistólica, Pressão Diastólica.

Após definição das variáveis de entrada da rede, a etapa seguinte foi a definição do conjunto de treino e de teste. Primeiramente, a amostra foi dividida em dois conjuntos distintos: 80% da amostra para compor o conjunto de treino (1.651 adolescentes) e o restante foi alocado para teste (413 adolescentes).

Considerando a prevalência de SM na amostra total (10,2%), apenas de 211 adolescentes possuem SM. Assim, tomando 80% da amostra total para o conjunto de treino, apenas 170 adolescentes possuíam SM nesse conjunto. Dessa forma, para evitar baixa representação de SM no conjunto de treino, foi realizado o balanceamento dos dados e em seguida a divisão da amostral total (80% treino e 20% teste). Para fins de comparação, também foram treinadas redes considerando os dados desbalanceados. Para o balanceamento dos dados foi utilizado o algoritmo de ROSE por meio do pacote ROSE (LUNARDON; MENARDI; TORELLI, 2015) do R

Com relação a configuração da RNA, foram implementadas diferentes redes *feed-forward*, todas elas contendo apenas uma camada oculta, diferenciando apenas em relação ao número de neurônios da camada de entrada e na camada oculta. Para treinamento da rede foi utilizado o algoritmo *back-propagation* com no máximo 10.000 iterações, taxa de aprendizado de 0.25 e três diferentes parâmetro de decaímento (0.01, 0.1 e 0.5) para evitar o *overfitting* da rede. A função de ativação utilizada na camada oculta e de saída foi a função logística sigmoidal.

3.4 Avaliação de Desempenho da RNA

Para avaliar os resultados obtidos através das Redes Neurais Artificiais, utilizaremos cinco métricas de avaliação presentes na literatura atual quando se trata de avaliação de performance de uma rede neural.

Através de uma matriz de confusão, obtemos os seguintes valores: VP (Verdadeiros Positivos), VN (Verdadeiros Negativos), FP (Falsos Positivos), FN (Falsos Negativos).

As métricas utilizadas serão: Acurácia, Sensibilidade, Especificidade, Kappa e Curva ROC. A acurácia é a probabilidade da classificação estar correta, e é calculada da seguinte maneira:

$$Acurácia = \frac{VP + VN}{FN + FP + VP + VN} \tag{3.1}$$

A sensibilidade é a proporção de classes positivas que foram identificadas corretamente pela rede neural, e é calculada da seguinte maneira:

$$Sensibilidade = \frac{VP}{VP + FN} \tag{3.2}$$

A especificidade é a proporção de classes negativas que foram identificadas corretamente pela rede neural, e é calculada da seguinte maneira:

$$Especificidade = \frac{VN}{VN + FP} \tag{3.3}$$

O Kappa de Cohen (KRAEMER, 2015) é uma métrica utilizada na avaliação da aceitação de classificadores binários em métodos de aprendizado de máquina e é definido da seguinte forma:

$$\kappa = \frac{2 \cdot (TP \cdot TN - FP \cdot FN)}{(TP+FP)\cdot(FP+TN)+(TP+FN)\cdot(FN+TN)} \tag{2}$$
$$(\text{pior valor } = -1 \text{ ; melhor valor } = +1 \text{ })$$

A Curva ROC, ou *Reciever Operator Characteristic*, é uma métrica de avaliação para problemas de classificação binária. É uma curva probabilística que plota a Sensibilidade versus o FVP (Falso Valor Positivo) em vários valores de limiares e essencialmente separa o "sinal"do "ruído". A área debaixo da curva (AUC) é a medida da habilidade do classificador de distinguir entre classes e é utilizado como referência da curva ROC (BHANDARI, 2020).

Quanto maior a AUC, melhor a performance do modelo em distinguir entre classes positivas e negativas. Quado AUC = 1, o classificador será capaz de perfeitamente distinguir entre todos os pontos de classes negativas e positivas corretamente, caso seja 0, o classificador estará prevendo todos os negativos como positivos e vice-versa (BHANDARI, 2020).

Quando 0.5<AUC<1, tem uma chance alta do classificador ser capaz de distinguir os valores das classes positivas dos valores das classes negativas, isso porque o classificador

é capaz de detectar mais números de verdadeiros positivos e verdadeiros negativos do que falsos negativos e falsos positivos (BHANDARI, 2020).

4 Resultados

Neste capítulo são apresentados os resultados obtidos de acordo com a metodologia que foi proposta anteriormente. O principal objetivo é determinar a melhor rede neural artificial para predição de SM em adolescentes. A avaliação das redes neurais, como foi mostrado no capítulo anterior, foi feita através das métricas Acurácia, Sensibilidade, Especificidade, Kappa de Cohen e Curva ROC.

A amostra em estudo foi composta por 2.064 adolescentes com 18 e 19 anos de idade, sendo 50,9% do sexo feminino e IMC médio de 21,93 kg/m^2 e CC de 81,65 cm (Tabela 3). A prevalência de SM nas meninas foi de 7,7% e nos meninos 12%, sendo essa diferença estatisticamente significante (p<0,05).

Tabela 3 – Medidas Descritivas das Variáveis em estudo

Variáveis	Mínimo	Média	Máximo	Desvio Padrão
Idade (anos)	18	18	19	0,4
Glicemia (mg/dL)	11	92,23	356	16,02
HDL (mg/dL)	8	49,38	109	11,7
CC (cm)	58,95	81,65	136,9	9
Peso (Kg)	34,15	61,34	151,01	13,08
Altura (cm)	141	167	196	9,1
IMC (kg/m^2)	14,68	21,93	49,31	3,9
RCE	0,36	0,48	0,77	0,05
Pressão Sistólica (mmHg)	75	114,2	171	12,1
Pressão Diastólica (mmHg)	51	70	109	7,4
Triglicerídios (mg/dL)	8	90,8	672	48,5

A tabela 4 apresenta as medidas de desempenho para as redes neurais treinadas a partir dos dados originais (dados desbalanceados). Independentemente das variáveis de entrada da redes, todas as redes neurais treinadas apresentaram acurácia acima de 80%. Entretanto, apenas a RNA 1 e RNA 2 apresentaram o Kappa acima 60%, bem como boa sensibilidade e especificidade.

Tabela 4 – Medidas de desempenho da RNA para amostra de teste - Dados Desbalanceados

Rede Neural	Acurácia	Kappa	Sensibilidade	Especificidade
RNA 1	96.0%	80.0%	78.0%	98.0%
RNA 2	96.0%	81.0%	78.0%	98.0%
RNA 3	89.0%	27.0%	24.0%	97.0%
RNA 4	90.0%	28.0%	24.0%	97.0%
RNA 5	88.0%	30.0%	31.0%	95.0%
RNA 6	90.0%	26.0%	21.0%	97.0%
RNA 7	90.0%	27.0%	21.0%	98.0%

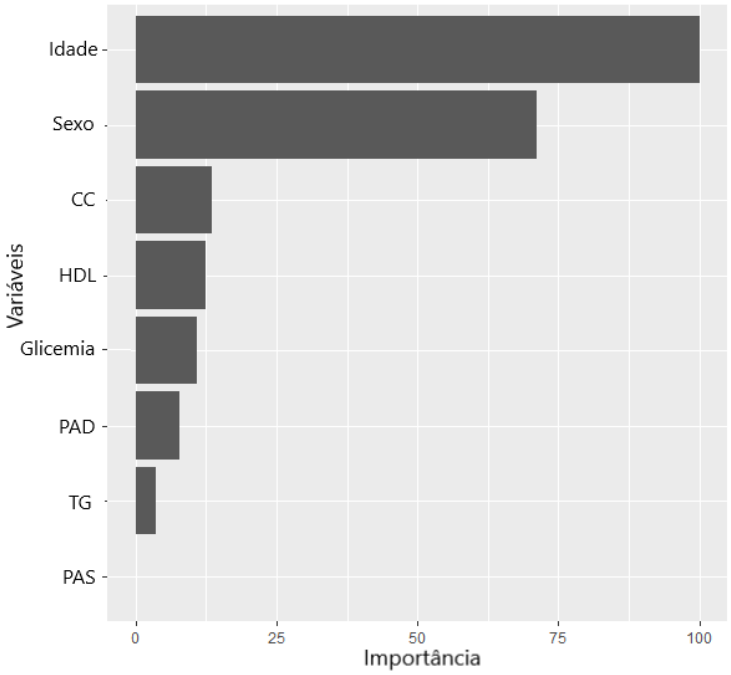
Quando considerados os dados balanceados, observou-se um aumento na sensibilidade (porcentagem de adolescentes com SM classificados corretamente) em todas as redes neurais avaliadas, tendo a RNA2 maior acurácia (83,0%), índice de Kappa (68,0%) e sensibilidade (85,0%), conforme apresentado na Tabela 5. Observa-se ainda que a RNA7, a qual não considera como variáveis de entrada os exames bioquímicos (glicemia, TG e HDL) apresentou boa acurácia (81,0%), sensibilidade (82,0%) e especificidade (81,0%), quando foi realizado o balanceamento dos dados (Tabela 5).

Tabela 5 – Medidas de desempenho da RNA para amostra de teste - Dados Balanceados

Rede Neural	Acurácia	Kappa	Sensibilidade	Especificidade
RNA 1	83.0%	66.0%	83.0%	83.0%
RNA 2	84.0%	68.0%	85.0%	83.0%
RNA 3	75.0%	50.0%	74.0%	76.0%
RNA 4	78.0%	57.0%	77.0%	79.0%
RNA 5	77.0%	53.0%	78.0%	75.0%
RNA 6	78.0%	57.0%	78.0%	79.0%
RNA 7	81.0%	63.0%	82.0%	81.0%

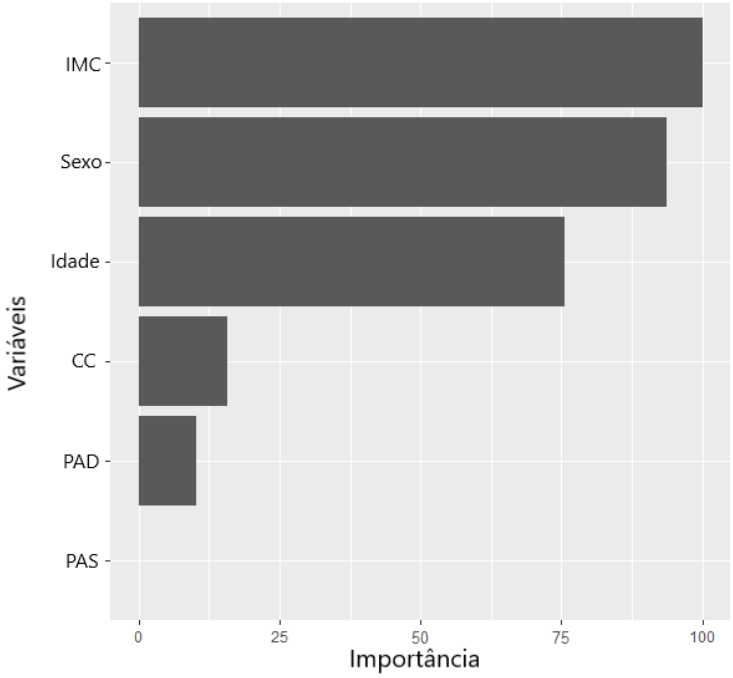
Considerando a RNA2 e RNA7 com os dados balanceados, foi avaliada a relevância das variáveis que compuseram os nós da camada de entrada de cada rede. Para a RNA2, tem-se que a idade e sexo apresentaram maior relevância. Além disso, as variáveis CC, HDL, glicemia e PAD apresentaram relevâncias próximas. Já as variáveis TG e PAS demonstraram baixa importância nesta rede para a determinação de SM (Figura 6).

Figura 6 – Importância das Variáveis - RNA 2 - Dados Balanceados



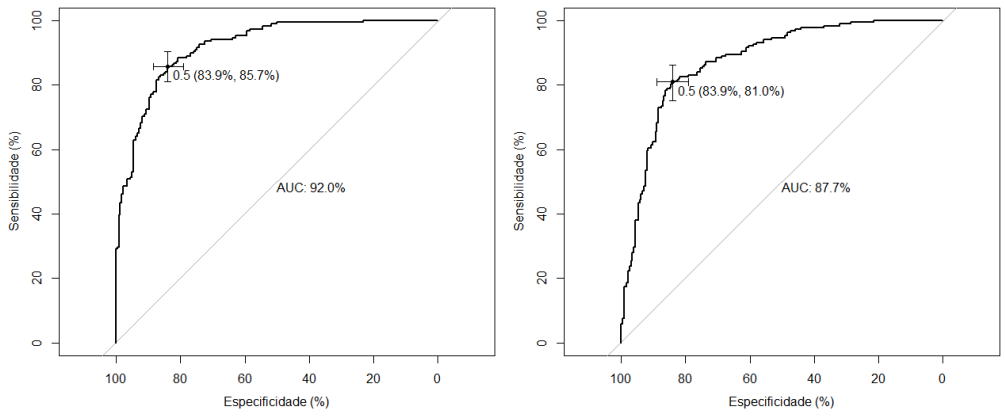
Na figura 7 são apresentadas a relevância das variáveis que compuseram a RNA7 com dados balanceados. Tem-se que o IMC foi a variável com maior importância, seguido das variáveis idade e sexo. Novamente, a PAS não apresentou relevância para o aprendizado da rede.

Figura 7 – Importância das Variáveis - RNA 7 - Dados Balanceados



Para avaliar o poder discriminatória da RNA2 e RNA7, foi construída a curva ROC. Ambas as redes apresentaram a AUC (Área sobre a curva) superior acima de 85,0% sendo considerado um desempenho satisfatório. Entretanto, a RNA2 teve o melhor desempenho com o AUC de 92,0%.

Figura 8 – Curvas ROC para as RNA treinadas com dados Balanceados



RNA2 RNA7

Na tabela 6, vemos uma comparação dos resultados obtidos neste trabalho com o que se encontra disponível na literatura. Haja visto que não existe nenhum trabalho publicado que relaciona o uso de RNA para predição de SM na amostra em estudo, compara-se com trabalhos que utilizaram a mesma técnica, na população adulta. Vemos que as redes obtidas neste estudo estão a par com as redes dos outros estudos, mostrando o potencial das RNA *feed-forward* para aprendizagem.

Tabela 6 – Comparação com os Trabalhos Relacionados

Trabalho	Acurácia	Sensibilidade	Especificidade	Curva ROC
Hirose et al. (2011)	-	93%	91%	-
Chen, Xiong e Ren (2014)	-	88%	83%	90%
Ivanović et al. (2016)	-	85%	83%	-
Este trabalho (RNA 2 - DB)	84%	83%	83%	92%
Este trabalho (RNA 7 - DB)	81%	82%	81%	87,7%

Legenda: DB - Dados Balanceados

5 Conclusão

Este trabalho teve como objetivo principal a utilização das RNA para predição de SM. E para obtermos um modelo estatístico capaz de identificar adolescentes com SM foram implementadas diferentes redes neurais artificiais. O desenvolvimento das redes consistiu inicialmente, a partir dos dados desbalanceados. Entretanto as redes treinadas não apresentaram resultados satisfatórios em relação a classificação dos adolescentes portadores de SM (sensibilidade).

Dessa forma, optou-se para o balanceamento dos dados, com a finalidade de obtermos um conjunto representativo do problema em estudo, bem como, uma estimativa não tendenciosa do erro de classificação dos modelos neurais. Os achados desse estudo mostram o bom potencial que o algoritmo ROSE possui para melhorar o aprendizado de classificadores sobre bases de dados desbalanceadas.

Com relação ao desempenho das redes neurais consideradas neste estudo, observamos, de um modo geral, uma equivalência das redes implementadas, ao analisarmos as acurácias sobre o conjunto de teste. Entretanto, elas diferem em relação à sensibilidade e especificidade, como também, em relação ao número de variáveis explicativas.

A rede alimentada com apenas as variáveis explicativas (RNA7) idade, sexo, IMC, CC, PAS e PAD constituí uma alternativa para a predição de SM em adolescentes. Esta rede, além de não necessitar dos exames bioquímicos do adolescente, apresentou uma boa acurácia, sensibilidade e especificidade quando comparada com a RNA que utiliza tais exames (RNA2). Além disso, tal rede foi capaz de identificar o IMC como variável relevante para a SM.

Apesar da SM ainda ser uma condição de difícil avaliação, sendo seus critérios questionados no mundo inteiro por diferentes especialistas, as RNA apresentadas neste estudo, podem ser utilizados como instrumentos para identificação precoce da SM. Sendo possível expandir suas possibilidades para o âmbito clínico por meio de aplicações utilizando diferentes populações de adolescentes.

O presente estudo mostrou a viabilidade das RNA na predição de SM. Sendo assim, surgem vários aspectos a serem considerados:

- Testar a viabilidade dos modelos propostos em populações com características sócio-epidemiológicas similares às características da amostra em estudo.
- Neste trabalho, utilizou-se apenas uma topologia de rede neural. Outras topologias de redes podem ser implementadas, entre elas, as Redes Neurais Convolucionais.

- Criação de um sistema neural de apoio ao diagnóstico precoce de SM, o qual utilizará o processamento neural. Uma interface fácil e eficiente poderá ser disponibilizada. Por meio do sistema neural, novas variáveis poderão ser incluídas para caracterização da SM, fornecendo dados que podem ajudar a uma tomada de decisão de com maior acurácia.

Referências

AGUDELO, G.; BEDOYA, G.; ESTRADA, A.; PATIñO, F.; MUñOZ, A.; VELáSQUEZ, C. Variations in the prevalence of metabolic syndrome in adolescents according to different criteria used for diagnosis: Which definition should be chosen for this age group? *Metabolic Syndrome and Related Disorders*, Apr 2014. Disponível em: <https://www.liebertpub.com/doi/10.1089/met.2013.0127>. Citado na página 15.

ALBERTI, G.; ZIMMET, P.; KAUFMAN, F.; TAJIMA, N.; SILINK, M.; ARSLANIAN, M.; WONG, G.; BENNET, P.; SHAW, J.; CAPRIO, S. *The IDF consensus definition of the METABOLIC SYNDROME IN CHILDREN AND ADOLESCENTS*. [S.l.]: International Diabetes Federation, 2005. ISBN 2-930229-49-7. Citado 2 vezes nas páginas 13 e 27.

ALMEIDA, C. A. Nogueira-de; HIROSE, T. S.; ZORZO, R. A.; VILANOVA, K. C. M.; RIBAS-FILHO, D. Critério da associação brasileira de nutrologia para diagnóstico e tratamento da síndrome metabólica em crianças e adolescentes. *International Journal of Nutrology*, v. 13, n. 03, p. 054–068, Dec 2020. Disponível em: <https://www.thieme-connect.de/products/ejournals/abstract/10.1055/s-0040-1721663>. Citado na página 12.

ALPAYDIN, E. *Introduction to machine learning*. [S.l.]: MIT press, 2020. Citado na página 16.

BHANDARI, A. 2020. Disponível em: <https://www.analyticsvidhya.com/blog/2020/06/auc-roc-curve-machine-learning/>. Citado 2 vezes nas páginas 29 e 30.

BONACCORSO, G. *Machine learning algorithms*. [S.l.]: Packt Publishing Ltd, 2017. Citado na página 16.

BROWNLEE, J. *Train-Test Split for Evaluating Machine Learning Algorithms*. 2020. Disponível em: <https://machinelearningmastery.com/train-test-split-for-evaluating-machine-learning-algorithms/>. Citado na página 22.

CHEN, H.; XIONG, S.; REN, X. Evaluating the risk of metabolic syndrome based on an artificial intelligence model. *Abstract and Applied Analysis*, v. 2014, 05 2014. Citado 2 vezes nas páginas 13 e 35.

CIABURRO, G.; VENKATESWARAN, B. *Neural Networks with R: Smart models using CNN, RNN, deep learning, and artificial intelligence principles*. [S.l.]: Packt Publishing Ltd, 2017. Citado na página 17.

COOK, S.; WEITZMAN, M.; AUINGER, P.; NGUYEN, M.; DIETZ, W. H. Prevalence of a Metabolic Syndrome Phenotype in Adolescents: Findings From the Third National Health and Nutrition Examination Survey, 1988-1994. *Archives of Pediatrics and Adolescent Medicine*, v. 157, n. 8, p. 821–827, 08 2003. ISSN 1072-4710. Disponível em: <https://doi.org/10.1001/archpedi.157.8.821>. Citado na página 27.

DIAS, S. M. *Democratizando a Inteligência Artificial*. 2020. Disponível em: <https://www.serpro.gov.br/menu/noticias/noticias-2019/democratizando-a-inteligencia-artificial>. Citado na página 16.

FERRANTI, S. D. de; GAUVREAU, K.; LUDWIG, D. S.; NEUFELD, E. J.; NEWBURGER, J. W.; RIFAI, N. Prevalence of the metabolic syndrome in american adolescents. *Circulation*, v. 110, n. 16, p. 2494–2497, Oct 2004. Disponível em: <https://www.ahajournals.org/doi/10.1161/01.cir.0000145117.40114.c7>. Citado 2 vezes nas páginas 9 e 27.

GOLLEY, R. K.; MAGAREY, A. M.; STEINBECK, K. S.; BAUR, L. A.; DANIELS, L. A. Comparison of metabolic syndrome prevalence using six different definitions in overweight pre-pubertal children enrolled in a weight management study. *International Journal of Obesity*, v. 30, n. 5, p. 853–860, Jan 2006. Disponível em: <https://pubmed.ncbi.nlm.nih.gov/16404409/>. Citado na página 15.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. *Deep Learning*. [S.l.]: MIT Press, 2016. <http://www.deeplearningbook.org>. Citado 4 vezes nas páginas 18, 19, 20 e 21.

HENDRIX, I. *Figura 3: Rede neural feedforward*. ResearchGate, 2011. Disponível em: <https://www.researchgate.net/figure/Figura-3-Rede-neural-feedforward_fig3_276857953>. Citado na página 18.

HIROSE, H.; TAKAYAMA, T.; HOZAWA, S.; HIBI, T.; SAITO, I. Prediction of metabolic syndrome using artificial neural network system based on clinical data including insulin resistance index and serum adiponectin. *Computers in Biology and Medicine*, v. 41, n. 11, p. 1051–1056, 2011. ISSN 0010-4825. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0010482511001909>. Citado 2 vezes nas páginas 13 e 35.

IVANOVIć, D.; KUPUSINAC, A.; STOKIć, E.; DOROSLOVAčKI, R.; IVETIć, D. Ann prediction of metabolic syndrome: a complex puzzle that will be completed. *Journal of Medical Systems*, v. 40, n. 12, Oct 2016. Disponível em: <https://link.springer.com/article/10.1007/s10916-016-0601-7#Tab6>. Citado 2 vezes nas páginas 13 e 35.

KAKUDI, H. A.; LOO, C. K.; MOY, F. M. Diagnosis of metabolic syndrome using machine learning, statistical and risk quantification techniques: A systematic literature review. Jun 2020. Disponível em: <https://www.medrxiv.org/content/10.1101/2020.06.01.20119339v1>. Citado na página 12.

KRAEMER, H. C. Kappa coefficient. *Wiley StatsRef: Statistics Reference Online*, p. 1–4, Jun 2015. Disponível em: <https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118445112.stat00365.pub2>. Citado na página 29.

LAHERA, G. *Unbalanced Datasets and What To Do About Them - Strands Tech Corner - Medium*. Strands Tech Corner, 2019. Disponível em: <https://medium.com/strands-tech-corner/unbalanced-datasets-what-to-do-144e0552d9cd>. Citado na página 23.

LEDERER, J. *Activation Functions in Artificial Neural Networks: A Systematic Overview*. 2021. Disponível em: <https://arxiv.org/abs/2101.09957>. Citado na página 20.

LUNARDON, N.; MENARDI, G.; TORELLI, N. *Package "ROSE" Title ROSE: Random Over-Sampling Examples*. [s.n.], 2015. Disponível em: <https://cran.r-project.org/web/packages/ROSE/ROSE.pdf>. Citado 2 vezes nas páginas 24 e 28.

NIELSEN, M. A. *Neural Networks and Deep Learning*. Determination Press, 2015. Disponível em: <http://neuralnetworksanddeeplearning.com/chap2.html>. Citado na página 21.

PIñA-AGUERO, M. I.; ZALDIVAR-DELGADO, A.; SALAS-FERNáNDEZ, A.; MARTíNEZ-BASILA, A.; BERNABE-GARCIA, M.; MALDONADO-HERNáNDEZ, J. Optimal cut-off points of fasting and post-glucose stimulus surrogates of insulin resistance as predictors of metabolic syndrome in adolescents according to several definitions. *Journal of Clinical Research in Pediatric Endocrinology*, v. 10, n. 2, p. 139–146, May 2018. Disponível em: <https://pubmed.ncbi.nlm.nih.gov/29082896/>. Citado na página 15.

ROSINI, N.; MOURA, S. A. Z. O.; ROSINI, R. D.; MACHADO, M. J.; SILVA, E. L. d. Metabolic syndrome and importance of associated variables in children and adolescents in guabiruba - sc, brazil. *Arquivos Brasileiros de Cardiologia*, 2015. Citado na página 12.

RUMELHART, D. E.; HINTON, G. E.; WILLIAMS, R. J. Learning representations by back-propagating errors. *Nature*, v. 323, p. 533–536, 1986. Citado na página 21.

SANTOS, A. M. d.; SEIXAS, J. M. d.; PEREIRA, B. d. B.; MEDRONHO, R. d. A. Usando redes neurais artificiais e regressão logística na predição da hepatite a. *Revista Brasileira de Epidemiologia*, v. 8, n. 2, p. 117–126, Jun 2005. Disponível em: <https://www.scielo.br/j/rbepid/a/wpHxNfpjJz4k9VHBRrzgVfw/abstract/?lang=pt>. Citado na página 27.

SHARMA, S.; SHARMA, S.; ATHAIYA, A. Activation functions in neural networks. *towards data science*, v. 6, n. 12, p. 310–316, 2017. Citado na página 20.

SIMõES, V. M. F.; BATISTA, R. F. L.; ALVES, M. T. S. S. d. B. e.; RIBEIRO, C. C. C.; THOMAZ, E. B. A. F.; CARVALHO, C. A. d.; SILVA, A. A. M. d. Saúde dos adolescentes da coorte de nascimentos de são luís, maranhão, brasil, 1997/1998. *Cadernos de Saúde Pública*, v. 36, n. 7, 2020. Disponível em: <https://www.scielo.br/j/csp/a/GZr9h3bDKmytNtHgJP3NjMz/?lang=pt>. Citado na página 26.

VINICIUS, A. *Redes Neurais Artificiais - Anderson Vinicius - Medium*. Medium, 2017. Disponível em: <https://medium.com/@avinicius.adorno/redes-neurais-artificiais-418a34ea1a39>. Citado na página 17.