UNIVERSIDADE FEDERAL DO AMAPÁ
BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO
DEPARTAMENTO DE CIÊNCIAS EXATAS E TECNOLÓGICAS

UMA ABORDAGEM BASEADA EM REDES NEURAIS ARTIFICIAIS PARA DIAGNÓSTICO DE DIABETES

LUCAS FAURO DE ARAÚJO

Orientador: Thiago Pinheiro do Nascimento

MACAPÁ
JANEIRO DE 2021

LUCAS FAURO DE ARAÚJO

UMA ABORDAGEM BASEADA EM REDES NEURAIS ARTIFICIAIS PARA DIAGNÓSTICO DE DIABETES

Trabalho de conclusão de curso apresentado à Universidade Federal do Amapá como requisito parcial para obtenção do título de Bacharel em Ciência da Computação.

Orientador: Thiago Pinheiro do Nascimento

MACAPÁ
JANEIRO DE 2021

Dados Internacionais de Catalogação na Publicação (CIP)
Biblioteca Central da Universidade Federal do Amapá
Elaborada por Cristina Fernandes– CRB-2/1569

Araújo, Lucas Fauro de.
Uma abordagem baseada em redes neurais artificiais para diagnóstico de diabetes. / Lucas Fauro de Araújo; orientador, Thiago Pinheiro do Nascimento. – Macapá, 2021.
44 f.

Trabalho de conclusão de curso (Graduação) – Fundação Universidade Federal do Amapá, Coordenação do Curso de Bacharelado em Ciência da Computação.

1. Aprendizado de máquina. 2. Redes neurais artificiais. 3. Diabetes - Diagnóstico. I. Nascimento, Thiago Pinheiro do, orientador. II. Fundação Universidade Federal do Amapá. III. Título.
005.1 A663
CDD. 22 ed.

UNIVERSIDADE FEDERAL DO AMAPÁ
DEPARTAMENTO DE CIÊNCIAS EXATAS E TECNOLÓGICAS
COORDENAÇÃO DO CURSO DE CIÊNCIA DA COMPUTAÇÃO

ATA DE DEFESA DE TCC

Realizou-se no dia 13 de janeiro de 2021, às 15:00, via videoconferência pelo Google Meet, a defesa de TCC intitulado “UMA ABORDAGEM BASEADA EM REDES NEURAIS ARTIFICIAS PARA DIAGNÓSTICO DE DIABETES”, do discente LUCAS FAURO DE ARAÚJO. A Banca Examinadora foi composta pelo Prof. Me. THIAGO PINHEIRO DO NASCIMENTO, presidente da banca e orientador; Prof. Me. MARCO ANTÔNIO LEAL DA SILVA e Prof. Esp. ADEILDO TELLES DA SILVA, examinadores. Concluída da defesa, foram realizadas as arguições e comentários. Em seguida, procedeu-se o julgamento pelos membros da Banca Examinadora, tendo o trabalho sido APROVADO com NOTA 9.

E, para constar, eu, Prof. THIAGO PINHEIRO DO NASCIMENTO, orientador e presidente da Banca Examinadora, lavrei a presente ata que, após lida e achada conforme, foi assinada por mim e demais membros da Banca Examinadora.

Macapá, 15 de fevereiro de 2021

Thiago Pinheiro do Nascimento

THIAGO PINHEIRO DO NASCIMENTO

PROF. ME. MARCO ANTÔNIO LEAL DA SILVA

PROF. ESP. ADEILDO TELLES DA SILVA

Aos meus pais e aos meus irmãos.

Agradecimentos

Inicialmente, agradeço à Deus, pela vida e saúde. Também agradeço aos meus pais, Geraldo e Denise, por todo suporte ao meu desenvolvimento acadêmico e todo o incentivo para me instiga aos estudos e à busca pelo conhecimento.

Agradeço aos meus irmãos, que sempre me apoiaram e foram referências para mim, principalmente nos momentos em que tive dificuldade.

Ao meu orientador, Thiago Pinheiro, pela oportunidade, incentivo e paciência durante o desenvolvimento desse projeto de pesquisa.

Aos docentes do Curso de Ciência da Computação da UNIFAP, que transmitiram todo conhecimento necessário para que eu me tornasse um profissional melhor; e aos meus colegas de curso, pelos momentos vividos e experiências compartilhadas.

Agradeço também aos meus familiares e amigos, que sempre estiveram juntos a mim durante esta fase, contribuindo direta ou indiretamente para que esse dia chegasse.

"As oportunidades multiplicam-se à medida que são agarradas." – Sun Tzu

Resumo

A diabetes registra grandes incidências mundialmente e é responsável pela diminuição da qualidade de vida de muitas pessoas. Sabe-se que diagnóstico prévio dessa doença é bastante importante, pois possibilita cuidados imediatos e a prevenção de riscos. O diagnóstico e classificação da diabetes são motivo de discussões ao longo de várias décadas, sendo o diagnóstico considerada uma tarefa complexa, devido ao fato das poucas metodologias existentes se basearem em sua maioria apenas nas alterações glicêmicas no organismo do indivíduo. Portanto o presente trabalho propõe uma abordagem baseada em aprendizado de máquina para o diagnóstico de diabetes.

Palavras-chave: Aprendizado de Máquina. Redes Neurais Artificiais. Diabetes. Diagnóstico.

Abstract

Diabetes has great incidence worldwide and is responsible for the decrease in the quality of life of many people. It is known that previous diagnosis of this disease is very important, as it allows immediate care and risk prevention. The diagnosis and classification of diabetes are the subject of discussions over several decades, and the diagnosis is considered a complex task, due to the fact that the few existing methodologies are based mostly on glycemic changes in the individual's body. Therefore, the present work proposes an approach based on artificial neural networks for the diagnosis of diabetes.

Keywords: Machine Learning. Diabetes. Diagnosis. Artificial Neural Networks

Lista de Figuras

Lista de Tabelas

Sumário

1 Introdução

A diabetes é uma epidemia que registra grandes incidências, sendo responsável pela diminuição da qualidade de vida de muitas pessoas. Sabendo que o diagnóstico da diabetes é considerado complexo e importante, o presente trabalho propõe a construção de um modelo baseado em aprendizado de máquina e extração de padrões, capaz de suportar esse diagnóstico. Para tanto, esse capítulo introdutório discute aspectos sobre o presente trabalho, tais como: a problemática a ser tratada, os objetivos, a hipótese, as justificativas, a metodologia a priori e as atividades a serem realizadas. O cronograma e a organização desse projeto de pesquisa também são apresentados ao final do capítulo.

1.1 Problemática

A diabetes se configura como uma das maiores epidemias do século XXI devido a sua crescente incidência (World Health Organization., 2016). A prevalência desta doença dobrou nos últimos 30 anos e significou um aumento de 4,7% para 8,5%. Estima-se que essa doença atinja 422 milhões de adultos atualmente, o que é considerado alarmante.

Hoje em dia a diabetes é considerada uma das principais síndromes de evolução crônica que acometem o homem (OLIVEIRA, J. E. P., 2004). É também uma doença associada a maiores taxas de hospitalização, maior utilização dos serviços de saúde como também responsável pela incidência de doenças cardiovasculares e cerebrovasculares, cegueira, insuficiência renal e amputação não traumática de membros inferiores (OLIVEIRA, 2017).

Em campanhas de rastreamento realizadas no Brasil em 2001, foi verificado que 50% da população diagnosticada não sabia que havia desenvolvido a doença (BARBOSA et al., 2001) e de acordo com Mazzini et al. (2013), esta doença é a sexta causa mais frequente de internações hospitalares.

Segundo o Ministério da Saúde (2006), são 4 milhões de mortes por ano relativas ao diabetes e suas complicações o que chega a representar 9% da mortalidade no mundo e um dos grandes impactos causados por esta doença está relacionado principalmente aos crescentes custos do tratamento da doença e suas complicações.

Em 1995, estimava-se que 4% da população adulta mundial estava acometida pela diabetes e que em 2025 esses chegariam a 5,4% da população. No Brasil, ao final da década de 80, era estimado que cerca de 8% da população entre 30 e 69 anos de idade estava acometida pela diabetes (Ministério da Saúde, 2006).

De acordo com MILECH, A., et al. (2016) a diabetes trata-se de uma doença

crônica não transmissível e um pequeno aumento em sua incidência é capaz de trazer potenciais implicações na saúde da população, e está caminhando para uma epidemia.

Em estudo realizado sobre rastreamento do risco do desenvolvimento de diabetes mellitus em pais de estudantes de uma escola privada na cidade de Jundiaí-SP, os resultados demonstraram que que os fatores de risco para o desenvolvimento do diabetes a maioria são modificáveis e que ações preventivas seriam importantes, além de que existem fatores de risco diferentes entre os sexos feminino e masculino, onde homens apresentam mais fatores do que as mulheres (MAZZINI et al., 2013).

Segundo o Ministério da Saúde (2006) a população com diabetes tipo 2 não sabe que é portadora desta doença e só procuram atendimento médico nas unidades básicas de saúde quando já manifestam sinais de complicação, por isto, testes de rastreamento são indicados para indivíduos assintomáticos com maior risco da doença. Alguns fatores são indicativos de maior risco para desenvolvimento da doença, tais como: idade acima de 45 anos, sobrepeso, obesidade central, antecedências, hipertensão, dentre outros (OLIVEIRA, 2017).

A classificação e o diagnóstico da diabetes são tarefas complexas que foram objeto de muito debate, consultas e revisão ao logo de várias décadas. Para diagnosticar a diabetes, há uma metodologia da World Health Organization (WHO) cujo critério para diagnóstico é através da observação dos níveis elevados de glicose no sangue, o que é considerado invasivo (International Diabetes Federation, 2017). Portanto, o presente trabalho propõe uma abordagem baseada em extração de padrões e aprendizado de máquina para suportar o diagnóstico de diabetes tipo 2, sendo possível a exploração dos resultados de variáveis não invasivas.

1.2 Objetivo Geral

O presente trabalho tem o objetivo de desenvolver uma abordagem baseada em aprendizado de máquina capaz de suportar o diagnóstico de diabetes tipo 2. É essencial ressalvar que essa abordagem não visa ser crucial para o diagnóstico de diabetes tipo 2, mas sim uma ferramenta para o apoio às tomadas de decisões médicas.

1.3 Objetivos Específicos

- Entender possíveis variáveis que levam o diagnóstico de diabetes tipo 2;
- Analisar bases de dados reais relacionadas ao diagnóstico de diabetes tipo 2;
- Extrair padrões das bases de dados sobre o diagnóstico de diabetes tipo 2;

- Aferir os resultados provenientes do reconhecimento de padrões aqui proposto;
- Comparar medidas de efetividade do estudo proposto com trabalhos relacionados.

1.4 Hipótese

A inteligência computacional é tradicionalmente utilizado em diversos contextos e exerce tarefas relacionadas ao raciocínio e ao aprendizado (RUSSELL et al., 2009). Nesse sentido, é comum observar casos de sucessos envolvendo a utilização da inteligência computacional em áreas da saúde, como o diagnóstico de doenças (Fuse et al., 2018; GRIF; AVUSH, 2018; Ahmad et al., 2017). Portanto, o presente trabalho parte do pressuposto que a inteligência computacional propõe abordagens atrativas no contexto do diagnóstico da diabetes, como o aprendizado de máquina.

1.5 Justificativa

A diabetes é um risco de nível mundial. Apesar dessa realidade, os métodos de diagnóstico desta doença são baseados apenas em observações dos níveis elevados de glicose no sangue, caracterizando-os como métodos invasivos.

O diagnóstico precoce da diabetes é essencial, uma vez que possibilita na redução e na prevenção de agravos vasculares para o paciente. Nesse contexto, é importante destacar que outros métodos fora a análise de níveis de hemoglobina glicada, também estão sendo testados. Contudo, os custos desses métodos ainda são elevados e, portanto é necessário que outros critérios sejam considerados (MAGALHÃES et al., 2011).

Métodos computacionais baseados em inteligência artificial já são utilizados no meio científico para auxiliar no processo de diagnóstico de doenças (KOUROU et al., 2015). Há evidências em que o diagnóstico de câncer de pele realizado pelos modelos computacionais tiveram resultados mais positivos quando comparado com aos de médicos ao utilizarem apenas os métodos tradicionais de diagnóstico (HAENSSLE et al., 2018).

Com relação a utilização de tais métodos para realizar o diagnóstico de diabetes, existem pesquisas em que se obteve taxa de sucesso em classificar os pacientes diabéticos e não-diabéticos acima dos 80%, o que já é um grande resultado que pode servir de auxílio para o diagnóstico precoce dessa doença. Em uma dessas pesquisas utilizou-se a abordagem de Máquina de Vetor de Suporte para realizar o aprendizado computacional sobre os dados, obtendo uma acurácia de 82% (KARATSIOLIS; SCHIZAS, 2012).

Portanto, o presente trabalho é atrativo pois propõe um método capaz de auxiliar o diagnóstico da diabetes através de características que geralmente o médico não

observa em métodos tradicionais.

1.6 Contribuição Científica

O trabalho proposto pretende gerar duas contribuições científicas, a se saber:

- Um abordagem capaz de diagnosticar diabetes tipo 2, explorando a utilização de variáveis que envolvam métodos invasivos e não invasivos;
- Um modelo capaz de prover apoio a tomada de decisões para profissionais médicos e de outras especialidades das áreas da saúde.

1.7 Cronograma e Atividades

O presente trabalho é pautado na realização das atividades listadas, como segue:

1. Revisão bibliográfica;
2. Investigação dos objetivos;
3. Identificação de recursos técnicos;
4. Implementação e avaliação de um estudo de caso;
5. Redação do projeto de pesquisa final para a conclusão de curso.
6. Apresentação final do trabalho de pesquisa para a conclusão do curso.

As atividades acima são efetuadas conforme o cronograma descrito na Tabela 1.

Tabela 1 – Cronograma

Etapa	abr.	mai.	jun.	jul.	ago.	set.	out.	nov.	dez.	jan.
1	x	x	x	x	x	x	x	x	x	
2	x	x								
3		x	x	x	x					
4					x	x	x			
5							x	x	x	
6										x

1.8 Organização

Além desse capítulo introdutório, este trabalho terá outros quatro capítulos. O segundo capítulo objetivará discutir sobre o referencial teórico relacionado ao trabalho proposto; o terceiro capítulo detalhará a metodologia a ser desenvolvida; o quarto visará a apresentação dos resultados; e, por fim, o quinto capítulo descreverá as conclusões.

2 Referencial Teórico

2.1 Diabetes Mellitus

De acordo com Ministério da Saúde (2006) a diabetes mellitus pode ser conceituada como:

> "Um grupo de doenças metabólicas caracterizada por hiperglicemias e associadas a complicações, disfunções e insuficiências de vários órgãos, especialmente olhos, rins, nervos, cérebro, coração e vasos sanguíneos. Pode resultar de defeitos de secreção e/ou ação da insulina envolvendo processos patogênicos específicos, por exemplo, destruição das células beta do pâncreas (produtoras de insulina), resistência à ação da insulina, distúrbios da secreção da insulina, entre outros. " (p. 9)

Atualmente a diabetes possui quatro classificações clínicas: Diabetes mellitus tipo 1, Diabetes Mellitus tipo 2, outros tipos específicos e a diabetes mellitus gestacional, onde estas são classificadas pela etiologia e não pelo tratamento, como é proposto pela Organização Mundial da Saúde (OMS).

A diabetes tipo 1 tem ocorrência de 5 -10% no total dos casos, onde neste tipo de diabetes ocorre a destruição das células β do pâncreas. É uma doença que ocorre principalmente em crianças e adolescentes, nos adultos ela pode também ocorrer, mas em qualquer um dos casos ocorre a destruição progressiva das células que geralmente leva a deficiência absoluta de insulina endógena (ASSOCIATION et al., 2014). O termo tipo I vai indicar destruição das células beta que eventualmente leva à um estágio de deficiência absoluta de insulina, fazendo com que o paciente necessite administrar doses de insulina para prevenir cetoacidose, coma e até mesmo a morte (Ministério da Saúde, 2006).

A diabetes do tipo 2 é correspondente a cerca de 90-95% dos casos, é o mais comum nos adultos, é um tipo de diabetes que está associado à histórico familiar, envelhecimento e estilo de vida pouco saudável. Está caracterizada pela resistência periférica à insulina, assim tenho aumento na produção de glicose pelo fígado o que acaba ocorrendo por conta das alterações na secreção pancreática de insulina (INZUCCHI et al., 2012).

A diabetes gestacional nada mais é do que a hiperglicemia diagnosticada no período gestacional, há uma intensidade variada, onde está se resolve principalmente no período pós-parto. De acordo com a OMS é recomendado detectar este tipo de diabetes com os mesmo procedimentos de diagnóstico empregados fora da gravidez, considerando como diabetes gestacional valores referidos fora da gravidez como indicativos de diabetes ou de tolerância à glicose diminuída (Ministério da Saúde, 2006).

Em relação aos outros tipos de diabetes podem ser citados os defeitos genéticos na função das células β, designados por MODY (Matury Onset Diabetes of The Young), na ação da insulina, doenças do pâncreas exócrino, diabetes induzida por fármacos entre outros. (PAIVA, 2001)

2.1.1 Diagnóstico da Diabetes (Métodos Tradicionais)

O diagnóstico da diabetes baseia-se nas alterações da glicose no organismo do indivíduo, existem dois métodos de verificação dessas alterações podendo ser pela medição da glicose plasmática em jejum de 8 horas, ou verificação após 2 horas de uma sobrecarga de glicose (75g) por via oral (GROSS et al., 2002). Podemos observar a relação destas medições com o diagnóstico diabetes na tabela 2.

Categoria	Jejum	TOTG - 75g - 2h	Casual
Normal	$<$110	$<$140	
Glicose plasmática de jejum alterada	$\geq$110 e $<$126		
Tolerância à glicose diminuída	$<$126	$\geq$140 e $<$200	
Diabetes mellitus	$\geq$126	$\geq$200	$\geq$200 com sintomas
Diabetes Gestacional	$\geq$110	$\geq$140	

Tabela 2 – Diagnóstico Diabetes Mellitus

2.1.2 Diagnóstico de Diabetes (Métodos Computacionais)

Apesar da convenção que é adotada pelo método descrito anteriormente, diversos fatores além da medição da glicose deve ser analisada para diagnosticar o paciente com diabetes, e isto acaba por dificultar o trabalho do médico. As avaliações dos dados obtidos de pacientes e as decisão que este profissionais tomam são fatores importantes no diagnóstico da diabetes. Para ajudar estes especialistas e mitigar erros que podem ser cometidos, os sistemas de classificação conseguem fornecer dados aos médicos para serem examinados em menor tempo e mais detalhados. Sistemas especialistas e técnicas de inteligência artificial utilizados na construção modelos de classificação em diagnóstico médico estão aumentando gradualmente (TEMURTAS; YUMUSAK; TEMURTAS, 2009).

As Redes Neurais Artificiais podem fornecer ferramentas importantes para ajudar os profissionais na interpretação de dados clínicos complexos. A maioria das aplicações de RNA utilizadas no ramo da medicina, resolvem tarefas de classifica-

ção, interpretando as características informadas e designando o paciente a uma classe específica (AL-SHAYEA, 2011).

Conforme exposto por Amato et al. (2013), existem diversas aplicações do uso de métodos computacionais em diagnósticos médicos dentro da literatura. Doenças cardiovasculares, câncer e diabetes, são classificadas como doenças mais graves, portanto o desenvolvimento de ferramentas para o diagnóstico destas doenças é de grande relevância. O quantitativo de dados clínicos destas é amplo, facilitando o desenvolvimento de modelos que possam auxiliar no diagnóstico médico de tais doenças.

2.2 Aprendizado de Máquina

Aprendizado de máquina é um ramo da inteligência artificial, que se dedica a resolver o problema de como os algoritmos podem aprender a partir de uma amostra de dados (MITCHELL, 2006). É uma técnica que tem como base a construção de modelos computacionais compostos por múltiplas camadas de processamento que aprendem a representar dados em vários níveis de abstração. Esse método melhorou significativamente tarefas como reconhecimento de voz, visão computacional e classificação de dados(LECUN; BENGIO; HINTON, 2015).

Dentro do campo de aprendizado de máquina podemos citar três principais abordagens que são utilizadas para a construção deste modelos, que são melhor explicados abaixo:

- Aprendizado Supervisionado: Neste aprendizado o conjunto de dados que temos acesso já possui os dados de entrada e as saídas desejadas.
- Aprendizado Não-Supervisionado: Neste aprendizado o valor de saída é desconhecido e os modelos são construídos para identificar padrões em grande conjunto de dados e agrupá-los.
- Aprendizagem por Reforço: É uma arquitetura onde não se sabe exatamente a saída correta do sistema, porém é possível avaliar o quão boa é para ponderar o processo de treinamento.

Ao aplicar uma abordagem de aprendizado de máquina, as amostras de dados representam os componentes básicos do sistema. Cada amostra desta possui um conjunto de características que é representada por diferentes tipos de valores. O objetivo principal de aplicar uma técnica dessas é produzir um modelo que possa ser usado para realizar classificação, predição ou tarefes semelhantes a estas. O processo de aprendi-

zagem de máquina é mais comumente utilizada para realizar tarefas de classificação (KOUROU et al., 2015).

Conforme exposto por Kourou et al. (2015), existem inúmeras técnicas e estratégias de pré - processamento de dados que visam modificar os dados para seu melhor aproveitamento em determinada abordagem para o aprendizado de máquina. Entre essas técnicas podemos citar: redução de dimensionalidade, seleção de atributos e extração de características. A técnica redução de dimensionalidade por exemplo consegue eliminar propriedades irrelevantes da base de dados, produzindo modelos mais robustos, além disto os algoritmos de ML tendem a trabalham melhor com dimensionalidades baixas.

2.3 Redes Neurais Artificiais

As redes neurais artificiais, é uma das abordagens de aprendizado de máquina, na qual muitos artigos identificam os seus benefícios em relação as técnicas dos modelos estatísticos tradicionais (GARDNER; DORLING, 1998). As redes neurais artificiais conseguem lidar com uma variedade de problemas de classificação ou reconhecimento de padrões. Elas são treinadas para gerar uma saída que seja a combinação entre os valores de entrada. Podem possuir múltiplas camadas escondidas, estas que representam matematicamente as conexões neurais que são utilizadas para construção destes modelos. Entretanto apesar de ser comumente utilizada, as redes neurais sofrem de algumas desvantagens, sua estrutura de camadas muitas vezes é demorada e possui um desempenho ruim. Além de que esta técnica é caracterizada como uma tecnologia de 'caixa preta'. Em outras palavras, tentar desvendar como o modelo executa o processo de classificação ou por que uma RNA não funcionou é quase impossível de ser detectado. (KOUROU et al., 2015)

2.3.1 Perceptron

Introduzido por Rosenblatt (1958), é inspirada na célula elementar do sistema nervoso central o neurônio. Perceptron é o algoritmo mais simples de aprendizado de máquina, é um modelo de aprendizado supervisionado para classificações binárias. Este algoritmo permite que os neurônios aprendam através de um conjunto base de treinamento, onde o algoritmo irá automaticamente definir os melhores pesos das ligações entre os neurônios.

Analisando a figura 1 podemos observar os elementos que compõem a arquitetura de uma rede perceptron. Da esquerda para a direita temos os sinais de entrada (X), os pesos sinápticos (w), o combinador linear ($\sum$) e por final a função de ativação (y). A

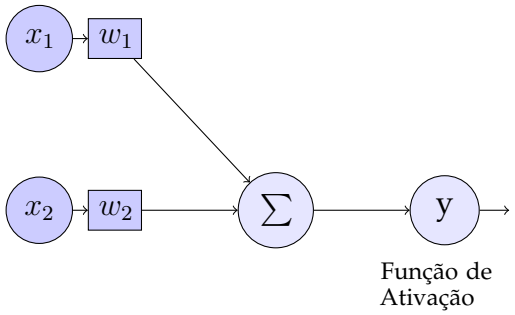

Figura 1 – Exemplo de arquitetura Perceptron

perceptron é tida como um classificador linear, podendo sua estrutura ser representada matematicamente da seguinte forma.

$$u = \sum_{i=0}^{n} w.x_i$$

$$y = g(u)$$

2.3.2 Multilayer Perceptron

A arquitetura Perceptron por ser simples consegue resolver apenas problemas que são linearmente separáveis como foi descrito na seção anterior. A solução para este problema é expandir o número de neurônios da arquitetura, adicionando 'camadas ocultas' ao modelo. Este tipo de abordagem ficou conhecido como Multilayer Perceptron (MLP). As redes neurais em geral são descritas por suas arquiteturas e seus algoritmos de aprendizagem, assim como o modelo simples de Perceptron, a MLP é uma técnica feed-forward, isto significa que as camadas são conectadas umas as outras em uma direção apenas, especificamente da camada de entrada a camada de saída. No caso da MLP além de ser feed-forward ela também é uma arquitetura fully-connected, ou seja esta estratégia significa que cada neurônio de uma camada está conectado a todos os neurônios da camada subsequente (POULTON, 2001).

A camada de entrada de uma MLP recebe os valores da base de dados de treinamento, o número de neurônios desta camada é fixo representado pelo total de padrões de treinamento existentes na base. A camada de saída pode ter valores variáveis dependendo do tipo de tarefa de classificação que está sendo realizada. Cada conexão (sinapse) entre os neurônios possui seu respectivo peso semelhante ao modelo Perceptron. O valor de saída de um neurônio se da pela pela soma da multiplicação dos valores

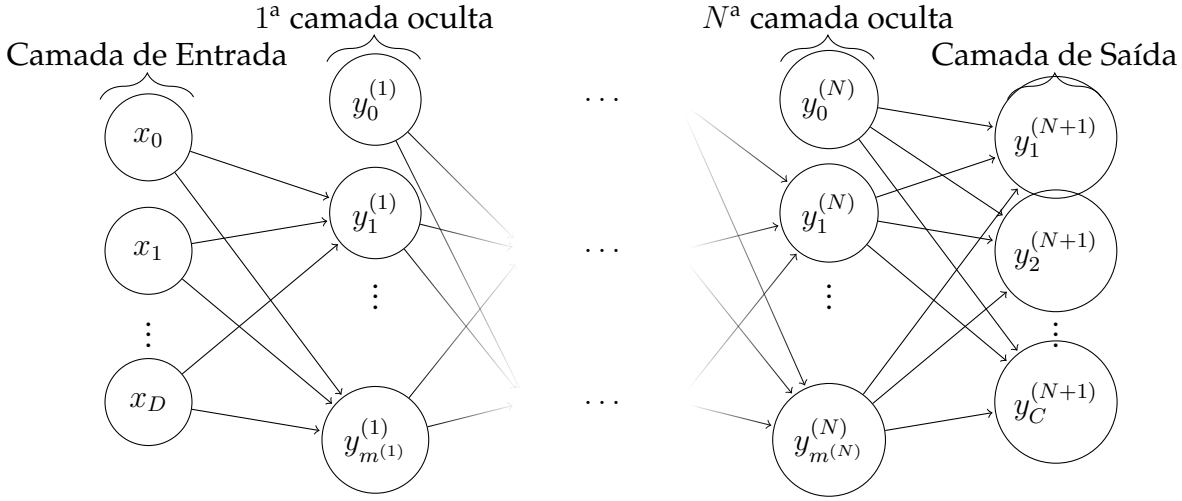

Figura 2 – Exemplo de uma rede neural Multilayer Perceptron

de entrada pelos seus respectivos pesos. Esta saída passa ainda por uma função de ativação para formar sinapse com a camada subsequente, este processo se repete até a saída de rede neural. Após a primeira interação os pesos sinápticos que são comumente inicializados de forma aleatória passam por um processo de atualização, este que é resultado da aplicação de um algoritmo de aprendizagem, um dos mais comuns é o **back-propagation** que realiza o ponderamento dos pesos de cada sinapse, realizando o aprendizado do modelo. (POULTON, 2001).

2.3.3 Funções de Ativação

Uma das unidades principais de redes neurais, as funções de ativação, são utilizadas para transformar o nível de ativação de um neurônio em um sinal de saída. Existem diversas funções de ativação que podem ser utilizadas nos modelos de Redes Neurais(KARLIK; OLGAC, 2011).

2.3.3.1 Função Sigmoide

A função sigmoide é definida por:

$$g(x) = \frac{1}{1 + e^{-x}}$$

Existem diversas vantagens em utilizar esta função no processo de treinamento das redes neurais que utilizam o algoritmo de back-propagation, pois ela consegue minimizar a capacidade computacional necessária para o treinamento (KARLIK; OLGAC, 2011).

2.3.3.2 Tangente Hiperbólica

Esta função é definida como a razão entre o seno hiperbólico e o cosseno hiperbólico, como segue abaixo:

$$\tanh(x) = \frac{\sinh(x)}{\cosh(x)} = \frac{e^x - e^{-x}}{e^x - e^{-x}}$$

2.3.3.3 RELU

A função relu possui um comportamento linear e é definida por:

$$ReLU(x) = \max\{0, x\}$$

2.3.3.4 Leaky ReLu

A função Leaky ReLu é semelhante a ReLu, porém ela permite pequenos valores para os resultados negativos :

$$LReLu(x) = \begin{cases} x & \text{se} \quad x \geq 0 \\ 0.01x & \text{se} \quad x < 0 \end{cases}$$

2.3.4 Backpropagation

Uma MLP é formada basicamente por três camadas, denominadas entrada, saída e pelo menos uma camada oculta. Backpropagation é uma técnica de aprendizado supervisionado utilizado para realizar o treinamento das redes neurais artificiais. O modo de aprendizado deste algoritmo é através do processamento interativo do conjunto de amostras de treino, comparando a previsão da rede para cada instância com seu respectivo rótulo da classe. Em outros termos, visando minimizar o erro entre a predição realizada pelo modelo e a sua classe real, os pesos sinápticos da rede são atualizados a cada amostra de treinamento (KAHRAMANLI; ALLAHVERDI, 2008). Em resumo o algoritmo executa as seguintes etapas:

1. A rede neural obtêm uma amostra de exemplo.

2. Compara a saída obtida da rede com a saída esperada pela a amostra. Realizando o calculo de erro para cada neurônio de saída.
3. Para cada neurônio restante da rede é realizado o cálculo do erro local.
4. Cada peso sináptico é ajustado visando diminuir o erro local.
5. Realiza as etapas anteriores para todas as amostras da base de treinamento.

3 Metodologia

3.1 Apresentação

Este trabalho propõe uma metodologia baseada em redes neurais artificiais, capaz de suportar a complexidade do diagnóstico de diabetes mellitus do tipo 2. O método proposto em resumo possui as seguintes etapas: aquisição dos dados, pré-processamento e classificação dos dados utilizando redes neurais artificiais. O fluxo-grama da figura 3, descreve as atividades a serem desenvolvidas.

A base de dados utilizada foi a dos Índios Pima consiste em 768 amostras, cada amostra possui um total de 8 (oito) variáveis clínica. Os valores desses atributos são de características clínicas oriundas de Mulheres de descendência dos Índios Pima com idade superior a 21 anos. (HAYASHI; YUKITA, 2016)

Como podemos observar na figura 5, esta base possui 500 amostras de pacientes não-diabéticos e 268 diabéticos, o que a torna desbalanceada. Esta característica se torna um problema para tarefa de classificação pois tende a valorizar a classe dominante no processo de treinamento.

As oito variáveis clínicas estão listadas abaixo:

- Número de vezes que a mulher engravidou
- Concentração de glicose no sangue (mg/dl)
- Pressão arterial (mmHg)
- Medida do triceps (mm)
- Quantidade de insulina em 2 horas de jejum (mu U/ml)
- IMC - Índice de Massa Corporal
- Ocorrências de casos de doença na família
- Idade em anos

O base de dados ainda possui um último atributo que é a classificação do paciente em diabéticos e não-diabéticos.

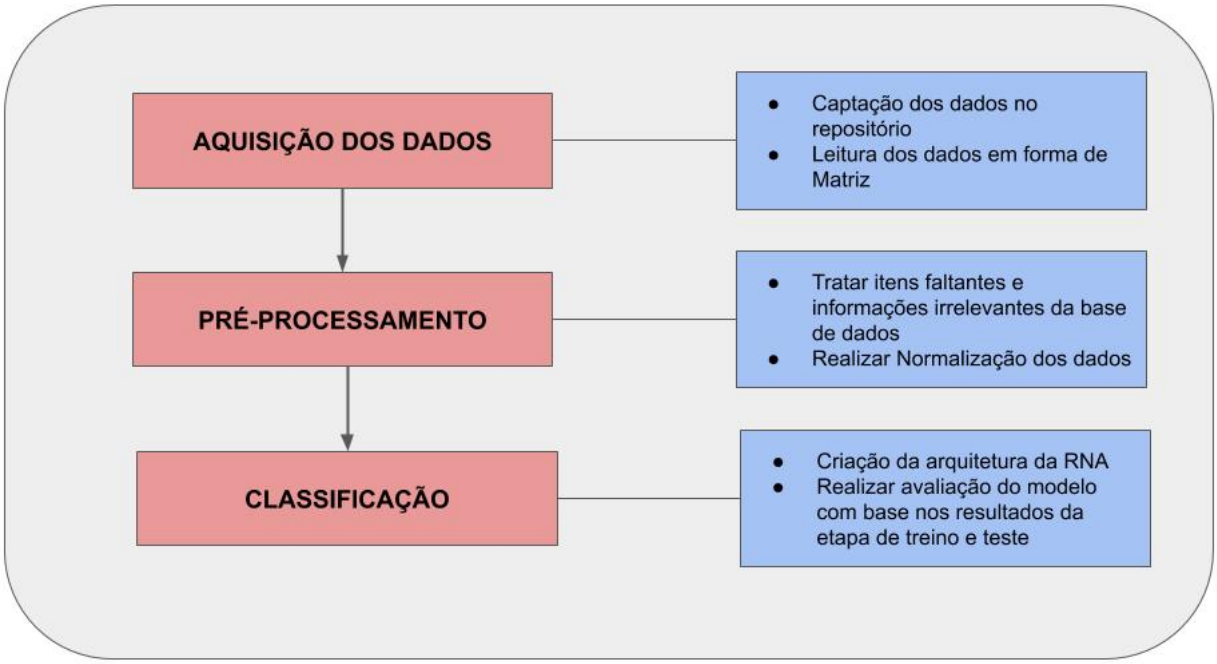

Figura 3 – Fluxograma do Trabalho Proposto

3.1.1 Ferramentas Utilizadas

O desenvolvimento do método proposto foi realizando utilizando algumas bibliotecas disponíveis, escritas na linguagem de programação Python. Através delas foi possível realizar o desenvolvimento das atividades de coleta e tratamento dos dados e criação do modelo de Redes Neurais Artificiais.

Os dados foram coletados usando a biblioteca de código aberto **pandas**, que possui ferramentas para auxiliar no processo de análise e manipulação de dados. Através dela foi possível realizar a visualização de forma detalhada das informações da base de dados, através de métodos existentes nesta biblioteca foi possível analisar de forma rápida e eficiente informações como: média dos atributos, fazer busca de dados faltantes e encontrar valores mínimos e máximos.

Na fase de pré-processamento e classificação utilizou-se o módulo scikit-learn, este é uma biblioteca de aprendizado de máquina de código aberto. Possui vários algoritmos que podem ser utilizado para realizar tarefas de classificação, além de outros que realizam funções de tratamento de dados (PEDREGOSA et al., 2011). Nesta pesquisa em questão o módulo foi utilizado para realizar as tarefas de normalização dos dados e busca exaustiva.

Para a construção dos modelos de redes neurais artificiais foi utilizado o TensorFlow juntamente com o Keras. TensorFlow é uma ferramenta de código aberto para a criação de modelos de aprendizado de maquina. Keras é uma API de alto nível que

roda sobre o tensorflow, com o foco no aprendizado profundo, esta ferramenta facilita a criação dos modelos pois, trás API's que podem ser utilizadas para a criação de redes neurais artificiais com maiores níveis de abstrações, facilitando o desenvolvimento destes modelos.

3.2 Aquisição dos Dados

Os dados estavam disponíveis e foram coletados através do repositório UCI-Irvine Machine Learning repository, após esta coleta os dados foram carregados e a sua transformação em matriz foi realizada com o auxílio do módulo pandas, descrito na seção anterior. Após a realização da coleta utilizou-se de métodos existentes na biblioteca para apresentar informações relevantes sobre a base de dados conforme demonstrado na figura 4, para que fosse realizado as análises iniciais da etapa de pré-processamento.

```
In [8]: data.describe()
```

Out[8]:

	num_preg	glucose_conc	diastolic_bp	skin_thickness	insulin	bmi	diab_pred	age
count	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000
mean	3.845052	120.894531	69.105469	20.536458	79.799479	31.992578	0.471876	33.240885
std	3.369578	31.972618	19.355807	15.952218	115.244002	7.884160	0.331329	11.760232
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.078000	21.000000
25%	1.000000	99.000000	62.000000	0.000000	0.000000	27.300000	0.243750	24.000000
50%	3.000000	117.000000	72.000000	23.000000	30.500000	32.000000	0.372500	29.000000
75%	6.000000	140.250000	80.000000	32.000000	127.250000	36.600000	0.626250	41.000000
max	17.000000	199.000000	122.000000	99.000000	846.000000	67.100000	2.420000	81.000000

Figura 4 – Descrição inicial da base de dados

3.3 Pré-Processamento dos dados

Esta etapa tem impacto significativo para o desempenho da rede neural, podendo melhorar a capacidade de generalização da mesma. As bases de dados podem ser carregadas de atributos sem que estes estejam relacionados ao objetivo real do modelo proposto. Portanto é necessário remover o máximo de informações irrelevantes e redundantes. Lidar com dados faltantes e realizar a normalização dos atributos também são tarefas realizadas durante a etapa de pré-processamento. (TEMURTAS; YUMUSAK; TEMURTAS, 2009)

Após a aquisição dos dados foi feito uma análise para verificar a qualidade das informações. Nesta análise buscou-se entender a relação dos atributos com o diagnóstico

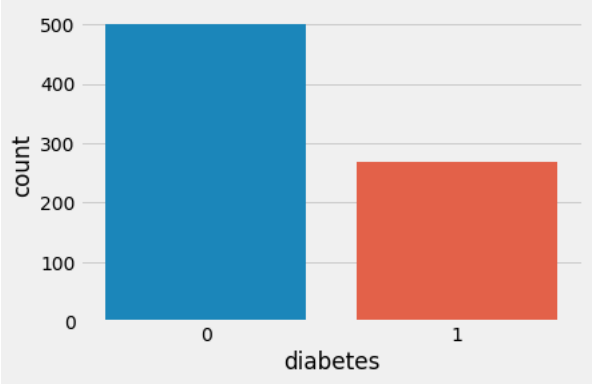

Figura 5 – Não-Diabéticos x Diabéticos

de diabetes mellitus, além de buscar por informações ausentes e/ou duplicados na base de dados.

Conforme descrito na seção anterior, a base de dados em questão é desbalanceada, sendo 500 pacientes não-diabéticos e 268 diabéticos. Pode-se observar na figura 6 que existem dados ausentes em algumas das instâncias. Para lidar com a tarefa de ausência de atributos podemos seguir algumas abordagens como:

- Ignorar instância com atributos desconhecidos
- Substituir os valores desconhecidos pelo valor mais comum da base
- Substituir os valores desconhecidos pela média dos valores existentes

Levando em consideração a base de dados em questão e o fato de existirem 374 instâncias sem a informação da insulina conforme demonstrado na figura 6, isso impossibilita a abordagem onde ignoramos os dados faltantes, pois isso comprometeria cerca de 50% da base de dados. A insulina e a medida do triceps tem valores pouco constantes, por isso, optou-se por utilizar a substituição pela média dos valores existentes.

```
glucose_conc       5
diastolic_bp      35
skin_thickness   227
insulin          374
bmi               11
dtype: int64
```

Figura 6 – Quantidade de dados faltantes

Para finalizar a etapa de pré-processamento, foi realizado a normalização dos valores das 8 variáveis clínicas de todas as instâncias presentes na base de dados. Para isto, utilizou-se a normalização min-max, que consiste em fazer o mapeamento das variáveis da seguinte maneira, para cada atributo o valor mínimo deste é transformado

em 0 e o máximo em 1, os restantes são convertidos para uma escala decimal entre 0 e 1 conforme demonstrado na figura 7, a fórmula desta conversão é apresentada na equação 1. Este procedimento é capaz de melhorar o desempenho da criação do modelo durante a fase de treinamento da rede neural.

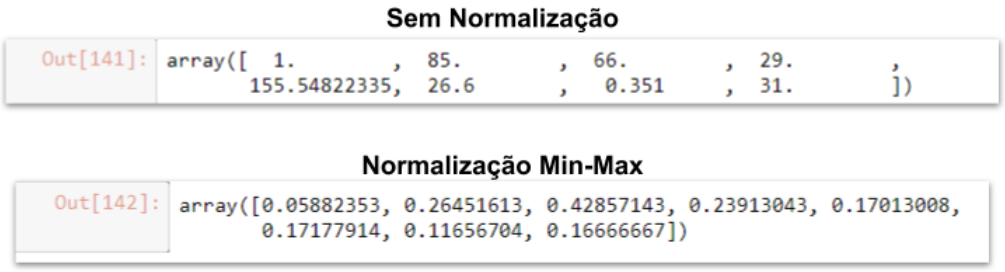

Figura 7 – Exemplo de Normalização Min-Max

$$x' = \frac{x - min(x)}{max(x) - min(x)} \tag{1}$$

3.3.1 Treino e Teste

A base foi dividida em duas partes, sendo elas denominadas de treino e teste. Realizamos a criação do modelo de classificação a partir dos dados contidos na base de treino e após isso o processo de avaliação deste modelo é feito utilizando a base de testes.

3.3.1.1 Split

Esta técnica se resume em dividir o total das instâncias em um nível de proporção, neste trabalho obteve-se melhores resultados utilizando uma divisão realizada em 75% para representar a base de treinamento e 25% dos dados restantes foram utilizados na fase de teste e validação.

3.4 Classificação

Esta fase consistiu em submeter as amostras de treino e teste extraídas no processo descrito na seção anterior a RNA para que esta realizasse a tarefa de classificação dos pacientes em diabéticos e não-diabéticos, conforme descrito na figura 8. Foi realizado a construção do modelo classificador utilizando redes neurais artificiais de arquitetura FeedForward de camadas múltiplas.

Diversos testes foram realizados com objetivo de se obter resultados melhores na fase de avaliação descrita na seção 3.5. Para isto inicialmente foi utilizado o módulo

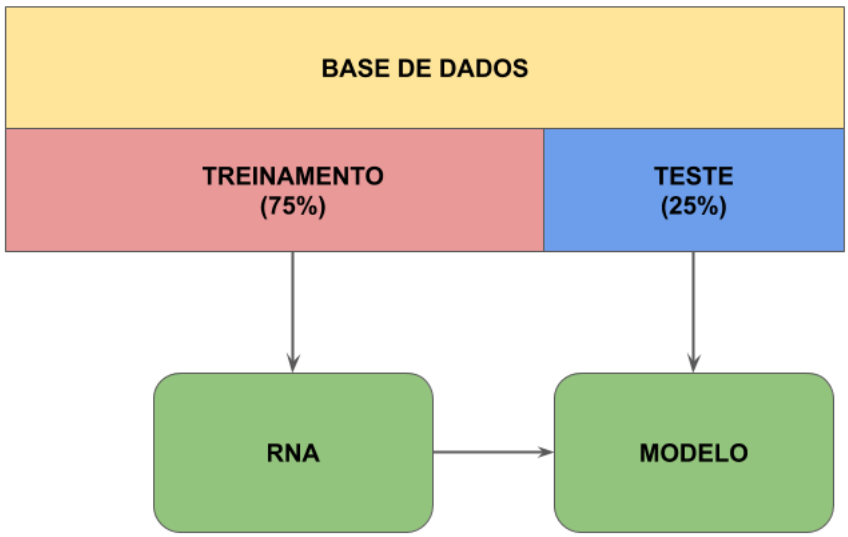

Figura 8 – Fluxo de desenvolvimento do classificador

scikit-learn através do método GridSearchCV, que é uma ferramenta que realiza busca exaustiva através de parâmetros especificados, visando identificar as melhores configurações para o modelo proposto, podemos observar na figura 9 alguns dos parâmetros utilizados neste processo. Optou-se por utilizar valores baixos para o número de épocas da etapa de treinamento da rede neural nesta busca, pois devido a quantidade de possibilidades, valores altos fariam com que este processo de treinamento tivesse um tempo de processamento muito elevado.

```
parametros = {'batch_size':[15,45,100],
              'epochs':[100,200,300],
              'optmizer': ['adam','sgd','rmsprop'],
              'loss':['hinge','mean_squared_logarithmic_error','logcosh','squared_hinge'
              'kernel_initializer': ['normal','random_uniform'],
              'activation':['softsign','tanh','softplus','relu'],
              'neurons':[15,30,40,100,300],
              'num_camadas':[1,2,3,4,5]
              }
```

Figura 9 – Parâmetros especificados no GridSearchCV

Após a realização dos testes iniciais chegamos a algumas arquiteturas que conseguiram alcançar resultados satisfatórios. Posteriormente a esta identificação, os melhores parâmetros apontados pelo GridSearchCV, foram identificados e utilizados na construção dos modelos seguintes. Estes passaram a ser construídos manualmente, e a partir deles foram realizados diversos testes com objetivo de identificar configurações que conseguissem alcançar resultados melhores.

Através dos parâmetros obtidos pela busca exaustiva foi possível identificar que a função de ativação Relu sempre esteve presente nas arquiteturas apontadas, por isto durante o processo de construção manual dos modelos ela foi escolhida para todos os testes, além disso sua variação Leaky Relu também foi utilizada durante este processo, objetivando avaliar o desempenho que esta teria na etapa de treinamento.

Podemos observar na tabela 3 os parâmetros utilizados na criação dos modelos de Redes Neurais Artificiais, nos testes 1 e 2 utilizou-se um número de camadas ocultas mais elevado, variando o valor de neurônios por camada oculta. Para os modelos criados nos testes 3, 4 e 5, o valor foi inferior e nos dois últimos teste optou-se por modificar a função de ativação.

Teste	Nº de Camadas Ocultas	Nº de Neurônios por Camada	Função de Ativação
1	5	15	Leaky Relu
2	5	30	Leaky Relu
3	3	40	Leaky Relu
4	3	30	Relu
5	3	300	Relu

Tabela 3 – Parâmetros das Redes Neurais Artificiais

No teste realizado de número 5, onde as camadas ocultas tiveram a presença de 300 neurônios, utilizou-se de uma técnica denominada dropout, esta técnica visa reduzir o overfitting. Algumas redes neurais tendem a construir adaptações que funcionam para os dados de treinamento porém não conseguem generalizar para os dados da base de teste. Esta técnica consiste em realizar a desativação dos neurônios da rede neural, para isto é atribuído a cada neurônio, uma probabilidade deste ser desativado durante a etapa de treinamento da rede neural. (SRIVASTAVA et al., 2014)

Os demais parâmetros de configuração da rede neural foram mantidos iguais para todos os testes finais realizados, estes parâmetros demonstraram resultados satisfatórios tanto na etapa de busca exaustiva quanto na construção dos modelos manualmente, foram eles:

- Função de Perda: Erro Quadrado Médio (MSE)
- Batch Size: 30
- Otimizador: Adam
- Épocas: 2000

A função de perda é responsável por calcular a diferença entre a previsão da rede neural e o resultado esperado, através dessa função a rede neural realiza os ajustes

nos pesos dos neurônios visando minimizar o resultado desta função, a equação 2 representa a fórmula da função do erro quadrado médio utilizada no desenvolvimento do modelo proposto.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2 \tag{2}$$

Batch Size representa o tamanho do lote em que o conjunto de dados de entrada será dividido em cada época do processo de treinamento. Esta divisão faz com que durante uma mesma época o modelo sofra mais de um ajustes nos seus parâmetros, ou seja, para cada lote que o conjunto de dados for divido, o algoritmo irá realizar o processo de ajustes dos pesos da rede neural. A utilização de valores pequenos ajuda a reduzir o custo de memória necessária para a realização do treinamento.

Os otimizadores são algoritmos responsáveis por realizar a tarefa de atualização dos parâmetros da rede neural. Épocas representa o número de interações que a rede neural irá realizar durante o processo de treinamento.

Além dos parâmetros supracitados, a camada de entrada da rede neural era composta por oito neurônios representando todas as variáveis clínicas da base da dados. A sua camada de saída era composta apenas por um neurônio e com a função de ativação sigmoid. A configuração da camada de saída é responsável por realizar a tarefa de classificação binária deste modelo.

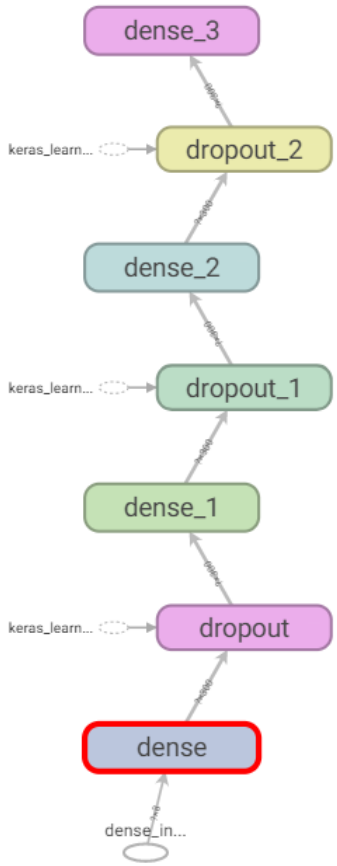

Figura 10 – Gráfico conceitual gerado pelo TensorFlow

A figura 10 representa um gráfico conceitual da estrutura de um modelo, este gráfico foi construído utilizando uma ferramenta existente dentro da biblioteca do TensorFlow. Além dos gráficos conceituais esta ferramenta consegue gerar dados de diagnóstico que foram utilizados durante o desenvolvimento da pesquisa para analisar o comportamento do modelo nos processos de treinamento e validação.

Como foi detalhado no início desta seção a base de dados utilizada para a construção do modelo é desbalanceada, com o objetivo de melhorar o desempenho da RNA, e para evitar que a mesma se adaptasse muito a classe dominante utilizou-se de uma técnica denominada SMOTE, que é uma abordagem que realiza o balanceamento da base de dados através da geração de exemplos sintéticos baseados nos vizinhos próximos de alguma instância existente na base de dados. Através do algoritmo utilizado é possível definir a variável K que representa o número de vizinhos a ser utilizado para a construção dos exemplos sintéticos (CHAWLA et al., 2002). Podemos observar na figura 11 o balanceamento realizado através do SMOTE para a base de dados utilizada.

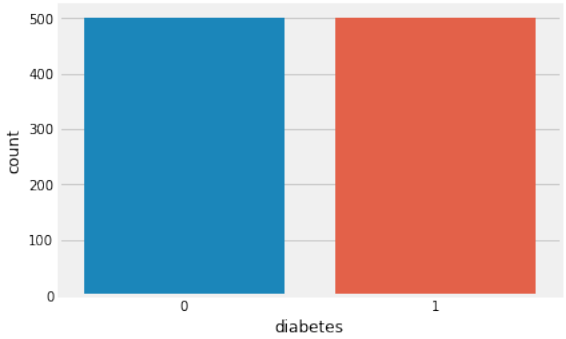

Figura 11 – Base de dados balanceada através do SMOTE

Após esse balanceamento, as arquiteturas definidas nos processos anteriores foram utilizadas para recriar os modelos e realizar o processo de treinamento utilizando a base de dados balanceada. Apesar deste esforço os resultados obtidos com a base de dados criada pelo o SMOTE não demonstrou resultados significativos, ficando próximos ou abaixo dos modelos criados anteriormente.

3.5 Avaliação

Esta etapa consiste em realizar a avaliação do classificador, para isto é construído uma matriz de confusão 2x2 contendo informações acerca dos erros e acertos do modelo treinado. Estas variáveis são comumente conhecidas como: Verdadeiro Positivo(VP), Falso Positivo(FP), Verdadeiro Negativo(VN) e Falso Negativo(FN); e são necessárias para a construção dos critérios de sensibilidade, especificidade e acurácia do classificador, que servem como métrica para avaliar sua capacidade de classificação. Estas variáveis são melhor explicadas abaixo:

- Verdadeiro Positivo(VP): Representa a quantidade de vezes que o modelo classifica corretamente o paciente diabético;
- Falso Positivo(FP): Representa a quantidade de vezes que o modelo classifica como diabético o paciente não-diabético;
- Verdadeiro Negativo(VN): Representa a quantidade de vezes que o modelo classifica corretamente o paciente não-diabético;
- Falso Negativo(FN): Representa a quantidade de vezes que o modelo classifica como não-diabético o paciente diabético.

A sensibilidade indica quão bom é o classificador para identificar os pacientes diabéticos e é definida por:

$$sensibilidade = \frac{VP}{(VP + FN)}$$

A especificidade indica quão bom é o classificador para identificar os pacientes não diabéticos e é definida por:

$$especificidade = \frac{VN}{(VN + FP)}$$

A acurácia é a taxa de sucesso ou acerto do teste e é calculada por:

$$acuracia = \frac{VP + VN}{(VP + VN + FP + FN)}$$

4 Resultados e Discussão

Através do método RNA utilizado neste estudo foi possível observar os seguintes resultados na tabela 4, para os 5 testes feitos na base de dados em questão. Com objetivo de avaliar a influência das variáveis no resultado, foram realizados 5 testes, utilizando os parâmetros informados na tabela 3, variando número de camadas, de neurônios e função de ativação, sendo observado ao final dos testes que quanto maior o número de neurônios utilizados, maior foi a acurácia. Possível observar também que o número de camadas é um outro fator que influenciou nos valores de acurácia, onde no teste 1 utilizando 5 camadas foi possível observar uma acurácia de 78% e no teste 3 apenas 4 camadas foram utilizadas apresentando apenas 70% de acurácia, mesmo que este apresentasse um número maior de neurônios.

Teste	Sensibilidade	Especificidade	Acurácia
Teste 1	0.61	0.87	0.78
Teste 2	0.65	0.7	0.69
Teste 3	0.65	0.73	0.7
Teste 4	0.64	0.73	0.7
Teste 5	0.7	0.89	0.828

Tabela 4 – Testes Realizados

Tendo em vista o trabalho em questão que visa o diagnóstico de diabetes, os resultados obtidos na métrica de sensibilidade são extremamente importantes, pois identificar os pacientes diabéticos são prioridades neste contexto. Podemos observar que nos teste realizados a sensibilidade nós trouxe valores abaixo do esperado em comparação com as outras métricas, estes resultados podem ser visto como um problema pois sugerem que o modelo criado não foi tão preciso na realização da tarefa proposta. Porém, nota-se uma evolução nos testes realizados, onde no teste de número 5, já foi possível alcançar sensibilidade na casa dos 70%. Como foi esclarecido nas seções anteriores, a base de dados em questão é desbalanceada, tendo menos instâncias de diabéticos, o que pode ter influenciado no modelo, que pode ter se adaptado melhor aos pacientes não-diabéticos.

4.1 Discussão

Neste estudo foi realizado teste de classificação seguindo utilização da base de dados completa, com as 8 características invasivas e não-invasivas, descritas no início

da seção 3, como feito por outros pesquisadores, desta forma foi possível observar uma acurácia de 82.8% utilizando o método de RNA para construção do modelo responsável pelo diagnóstico de diabetes mellitus tipo II, contudo é importante destacar que outros autores usando a mesma base dados, com métodos diferentes e/ou iguais obtiveram valores de acurácia abaixo do que foi encontrado nesta pesquisa, como o método SVM, onde é possível observar os valores de acurácia encontrado na tabela 5. Os testes realizados utilizando a mesma abordagem de redes neurais, observa-se valores entre 76% e 75% de acurácia, nosso método se difere das abordagens presentes na tabela pelo número de camadas e neurônios, pela a utilização de dropout e pela a alteração de alguns valores chaves na etapa de treinamento como a função de perda.

Tabela 5 – Comparação de resultados de diferentes classificadores

Método	Acurácia %	Autores
Logdisc	77.7	Statlog
IncNet	77.6	Norbert Jankowski
DIPOL92	77.6	Statlog
Linear Discr. Anal.	77.5-77.2	Statlog; Ster & Dobnikar
SVM, linear, C=0.01	77.5±4.2	WD-GM, 10XCV averaged 10x
SVM, Gauss, C, sigma opt	77.4±4.3	WD-GM, 10XCV averaged 10x
SMART	76.8	Statlog
GTO DT (5xCV)	76.8	Bennet and Blue
kNN, k=23, Manh, raw, W	76.7±4.0	WD-GM, feature weighting 3CV
kNN, k=1:25, Manh, raw	76.6±3.4	WD-GM, most cases k=23
ASI	76.6	Ster & Dobnikar
Fisher discr. analysis	76.5	Ster & Dobnikar
MLP+BP	76.4	Ster & Dobnikar
MLP+BP	75.8±6.2	Zarndt
LVQ	75.8	Ster & Dobnikar
LFC	75.8	Ster & Dobnikar
RBF	75.7	Statlog
NB	75.5-73.8	Ster & Dobnikar; Statlog
kNN, k=22, Manh	75.5	Karol Grudziński
MML	75.5±6.3	Zarndt
FSM stand. 5 feat.	75.4±4.9	WD, 10x10 test, CC>0.15
SNB	75.4	Ster & Dobnikar
BP	75.2	Statlog
SSV DT	75.0±3.6	WD-GM, SSV BS, node 5CV MC
kNN, k=18, Euclid, raw	74.8±4.8	WD-GM

CART DT	74.7±5.4	Zarndt
CART DT	74.5	Stalog
DB-CART	74.4	Shang & Breiman
ASR	74.3	Ster & Dobnikar
FSM standard	74.1±1.1	WD, 10x10 test
ODT, dyadic trees	74.0±2.3	Blanchard
Cluster means, 2 prototypes	73.7±3.7	MB
SSV DT	73.7±4.7	WD-GM, SSV BS, node 10CV strat
SFC, stacking filters	73.3±1.9	Porter
C4.5 DT	73.0	Stalog
C4.5 DT	72.7±6.6	Zarndt
Bayes	72.2±6.9	Zarndt
C4.5 (5xCV)	72.0	Bennet and Blue
CART	72.8	Ster & Dobnikar

Fonte: http://www.is.umk.pl/ duch/projects/projects/datasets.html

No estudo desenvolvido em Polat, Güneş e Arslan (2008), utilizando a classificação LS-SVM obteve uma sensibilidade (73.9%), especificidade (80%) e acurácia (78.2%) e para classificação a partir de GDA-LS-SVM, sensibilidade (73.9%), especificidade (82.05%) e acurácia de (79.16%), corroborando com os dados encontrados neste trabalho, ou seja faixa de acurácia entre 75-85%, mesmo que o método de classificação seja diferente. Os autores sugerem que a classificação e abordagem por aprendizado de maquina para diagnóstico de diabetes pode sim auxiliar no diagnóstico da doença e que este sistema pode ser útil para os médicos na decisão final sobre seus pacientes, ou seja utilizando ferramenta eficiente eles podem tomar decisões precisas

Ao analisar as variáveis foi possível observar que quando alteradas os resultados apresentavam resultados para acurácia, sensibilidade e especificidade variáveis para cada parâmetro alterado, de acordo com Ismailov (2014), definir quantos neurônios na camada oculta devem ser captados é necessário e que quanto maior o número desses neurônios, maior a probabilidade da rede fornece resultados precisos. Para Basheer e Hajmeer (2000), a determinação do número de camadas ocultas e de nós ocultos em cada camada é uma das tarefas mais críticas no design da RNA, ou seja, uma rede com muito poucos nós ocultos seria incapaz de diferenciar padrões complexos, contudo, redes com muitos nós ocultos seguirá ruídos nos dados devido ao excesso de parametrização, levando a uma generalização ruim dos dados não treinados.

Na pesquisa realizada foi possível observar que função de ativação relu com 3 camadas e 300 neurônios apresentou uma maior acurácia, além de sensibilidade e

especificidade quando comparado com a função Leaky ReLU, para Ertuğrul (2018) determinar uma função de ativação ideal em RNA é uma questão importante porque está diretamente ligada as taxas de sucesso obtido nos resultados, contudo a função de ativação ideal geralmente é determinada por testes ou ajustes.

Diante dos fatos analisados é notável que a MLP se torna eficiente para solução de problemas, como para o diagnóstico da diabetes mellitus II observado em nossa pesquisa, importante evidenciar que a MLP se torna cada vez mais atrativa por suas características notáveis de processamento de informação, pertinente principalmente à não linearidade, alto paralelismo, tolerância a falhas e ruídos e recursos de aprendizado e generalização (BASHEER; HAJMEER, 2000).

5 Conclusão

Tendo em vista os objetivos propostos para o trabalho em questão, podemos afirmar que foi possível desenvolver um modelo baseado em redes neurais artificiais capaz de suportar o diagnóstico de diabetes mellitus tipo 2.

Foi possível observar que a etapa de pré-processamento onde avaliamos a qualidade dos dados e tratamos eles para obter melhores resultados, bem como o design da RNA, são fatores importantes que refletem o desempenho final do modelo criado.

Na pesquisa realizada o teste que obteve melhor resultado alcançou uma acurácia de 82.28%, isto demonstra que o modelo é eficiente, podendo ser utilizado como auxílio no diagnóstico da doença.

Como foi relatado anteriormente, existem diversos exemplos na literatura de sistemas inteligentes auxiliando na tarefa de diagnóstico de doenças, porém estes modelos ainda devem ser visto apenas como ferramentas para apoiar as tomadas de decisões.

Desta forma, trabalhos futuros voltados para o desenvolvimento de modelos para o diagnósticos de diabetes podem ser realizados, a utilização de outra base de dados com um número maior de instâncias, mais balanceada e contendo um número maior de atributos podem favorecer a criação de um modelo baseado em redes neurais artificiais com uma capacidade de generalização maior do que a desenvolvida no trabalho em questão.

Referências

Ahmad, M. et al. Diagnostic decision support system of chronic kidney disease using support vector machine. In: *2017 Second International Conference on Informatics and Computing (ICIC)*. [S.l.: s.n.], 2017. p. 1–4.

AL-SHAYEA, Q. K. Artificial neural networks in medical diagnosis. *International Journal of Computer Science Issues*, Citeseer, v. 8, n. 2, p. 150–154, 2011.

AMATO, F. et al. *Artificial neural networks in medical diagnosis*. [S.l.]: Elsevier, 2013.

ASSOCIATION, A. D. et al. Diagnosis and classification of diabetes mellitus. *Diabetes care*, Am Diabetes Assoc, v. 37, n. Supplement 1, p. S81–S90, 2014.

BARBOSA et al. Campanha nacional de detecção de casos suspeitos de diabetes mellitus no brasil: relatório preliminar. *Revista Panamericana de Salud Pública*, SciELO Public Health, v. 10, p. 324–327, 2001.

BASHEER, I. A.; HAJMEER, M. Artificial neural networks: fundamentals, computing, design, and application. *Journal of microbiological methods*, Elsevier, v. 43, n. 1, p. 3–31, 2000.

CHAWLA, N. V. et al. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, v. 16, p. 321–357, 2002.

ERTUĞRUL, Ö. F. A novel type of activation function in artificial neural networks: Trained activation function. *Neural Networks*, Elsevier, v. 99, p. 148–157, 2018.

Fuse, H. et al. Detection of alzheimer's disease with shape analysis of mri images. In: *2018 Joint 10th International Conference on Soft Computing and Intelligent Systems (SCIS) and 19th International Symposium on Advanced Intelligent Systems (ISIS)*. [S.l.: s.n.], 2018. p. 1031–1034.

GARDNER, M. W.; DORLING, S. Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. *Atmospheric environment*, Elsevier, v. 32, n. 14-15, p. 2627–2636, 1998.

GRIF, M. G.; AVUSH, Y. The development of medical diagnostic system based on integration of traditional and eastern medicines. In: *XIV International Scientific-Technical Conference on Actual Problems of Electronics Instrument Engineering (APEIE)*. [S.l.: s.n.], 2018. p. 511–515. ISSN 2473-8573.

GROSS, J. L. et al. Diabetes melito: diagnóstico, classificação e avaliação do controle glicêmico. *Arquivos Brasileiros de Endocrinologia & Metabologia*, SciELO Brasil, v. 46, n. 1, p. 16–26, 2002.

HAENSSLE, H. et al. Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. *Annals of Oncology*, Oxford University Press, v. 29, n. 8, p. 1836–1842, 2018.

HAYASHI, Y.; YUKITA, S. Rule extraction using recursive-rule extraction algorithm with j48graft combined with sampling selection techniques for the diagnosis of type 2 diabetes mellitus in the pima indian dataset. *Informatics in Medicine Unlocked*, Elsevier, v. 2, p. 92–104, 2016.

International Diabetes Federation. *IDF Diabetes Atlas*. [S.l.]: International Diabetes Federation, 2017. v. 8.

INZUCCHI, S. E. et al. Management of hyperglycemia in type 2 diabetes: a patient-centered approach: position statement of the american diabetes association (ada) and the european association for the study of diabetes (easd). *Diabetes care*, Am Diabetes Assoc, v. 35, n. 6, p. 1364–1379, 2012.

ISMAILOV, V. E. On the approximation by neural networks with bounded number of neurons in hidden layers. *Journal of Mathematical Analysis and Applications*, Elsevier, v. 417, n. 2, p. 963–969, 2014.

KAHRAMANLI, H.; ALLAHVERDI, N. Design of a hybrid system for the diabetes and heart diseases. *Expert systems with applications*, Elsevier, v. 35, n. 1-2, p. 82–89, 2008.

KARATSIOLIS, S.; SCHIZAS, C. N. Region based support vector machine algorithm for medical diagnosis on pima indian diabetes dataset. In: IEEE. *2012 IEEE 12th International Conference on Bioinformatics & Bioengineering (BIBE)*. [S.l.], 2012. p. 139–144.

KARLIK, B.; OLGAC, A. V. Performance analysis of various activation functions in generalized mlp architectures of neural networks. *International Journal of Artificial Intelligence and Expert Systems*, v. 1, n. 4, p. 111–122, 2011.

KOUROU, K. et al. Machine learning applications in cancer prognosis and prediction. *Computational and structural biotechnology journal*, Elsevier, v. 13, p. 8–17, 2015.

LECUN, Y.; BENGIO, Y.; HINTON, G. Deep learning. *nature*, Nature Publishing Group, v. 521, n. 7553, p. 436, 2015.

MAGALHÃES, G. L. et al. Atualização dos critérios diagnósticos para diabetes mellitus utilizando a a1c. *HU Revista*, v. 37, n. 3, 2011.

MAZZINI, M. C. R. et al. Rastreamento do risco de desenvolvimento de diabetes mellitus em pais de estudantes de uma escola privada na cidade de jundiaí, são paulo. *Revista da Associação Médica Brasileira*, Elsevier, v. 59, n. 2, p. 136–142, 2013.

MILECH, A., et al. *Diretrizes da sociedade brasileira de diabetes*. Rio de Janeiro: [s.n.], 2016.

Ministério da Saúde. *Caderno de Atenção Básica - Diabetes Mellitus*. Brasília - DF: Ministério da Saúde, 2006.

MITCHELL, T. M. *The discipline of machine learning*. [S.l.]: Carnegie Mellon University, School of Computer Science, Machine Learning . . . , 2006. v. 9.

OLIVEIRA, e. a. Paulo de. *Diretrizes da Sociedade Brasileira de Diabetes*. [S.l.]: Sociedade Brasileira de Diabetes, 2017.

OLIVEIRA, J. E. P. Diabetes mellitus clínica, diagnóstico, tratamento multidisciplinar. In: *Conceito, Classificação e Diagnóstico do Diabetes.* [S.l.]: Atheneu, 2004. v. 1.

PAIVA, C. Novos critérios de diagnóstico e classificação da diabetes mellitus. *Medicina Interna*, v. 7, n. 4, p. 234–38, 2001.

PEDREGOSA, F. et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, v. 12, p. 2825–2830, 2011.

POLAT, K.; GÜNEŞ, S.; ARSLAN, A. A cascade learning system for classification of diabetes disease: Generalized discriminant analysis and least square support vector machine. *Expert systems with applications*, Elsevier, v. 34, n. 1, p. 482–487, 2008.

POULTON, M. M. Multi-layer perceptrons and back-propagation learning. In: *Handbook of Geophysical Exploration: Seismic Exploration*. [S.l.]: Elsevier, 2001. v. 30, p. 27–53.

ROSENBLATT, F. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, American Psychological Association, v. 65, n. 6, p. 386, 1958.

RUSSELL et al. *Artificial Intelligence: A Modern Approach*. 3rd. ed. Upper Saddle River, NJ, USA: Prentice Hall Press, 2009. ISBN 0136042597, 9780136042594.

SRIVASTAVA, N. et al. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, JMLR. org, v. 15, n. 1, p. 1929–1958, 2014.

TEMURTAS, H.; YUMUSAK, N.; TEMURTAS, F. A comparative study on diabetes disease diagnosis using neural networks. *Expert Systems with applications*, Elsevier, v. 36, n. 4, p. 8610–8615, 2009.

World Health Organization. *Global Report on Diabetes.* 2016. <http://apps.who.int/iris/bitstream/10665/204871/1/9789241565257eng.df>. [Online; Acessado em 07-Dezembro-2018].