

**CESED - CENTRO DE ENSINO SUPERIOR E DESENVOLVIMENTO  
UNIFACISA – CENTRO UNIVERSITÁRIO  
CURSO DE BACHARELADO EM SISTEMAS DE INFORMAÇÃO**

**DIEGO KAZADI KALUNDA**

**APLICANDO MODELOS DE APRENDIZAGEM DE MÁQUINA PARA PREVER  
HIPERTENSÃO ARTERIAL: UMA ANÁLISE COMPARATIVA**

**CAMPINA GRANDE - PB**

**2021**

DIEGO KAZADI KALUNDA

APLICANDO MODELOS DE APRENDIZAGEM DE MÁQUINA PARA PREVER  
HIPERTENSÃO ARTERIAL: UMA ANÁLISE COMPARATIVA

Trabalho de Conclusão de Curso  
apresentado como pré-requisito para a  
obtenção do título de Bacharel em  
Sistemas de Informação pelo Centro  
Universitário de Ciências Sociais  
Aplicadas.

Área de Concentração: Aprendizagem de  
Máquina.

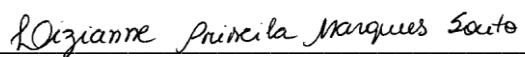
Orientadora: Profa. Msc. Lizianne Souto

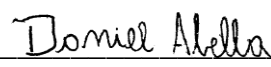
K14a	<p>Kalunda, Diego Kazadi.</p> <p>Aplicando modelos de Machine Learning para prever hipertensão arterial: uma análise comparativa. / Diego Kazadi Kalunda - Campina Grande-PB, 2021.</p> <p>Originalmente apresentada como Trabalho de Conclusão de Curso - Bacharelado em Sistemas de Informação do autor (Bacharel - UniFacisa - Centro Universitário, 2020).</p> <p>Referências.</p> <p>1. Aprendizagem de máquina. 2. Diagnóstico de hipertensão arterial. 3. Redes bayesianas. I. Título.</p> <p>CDU-004.7(043)</p>
------	-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

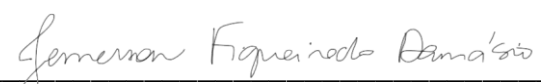
Trabalho de Conclusão de Curso,  
Aplicando modelos de *Machine Learning*  
para prever hipertensão arterial: uma  
análise comparativa, apresentado por  
Diego Kazadi Kalunda como parte dos  
requisitos para obtenção do título de  
Bacharel em Sistemas de Informação  
outorgado pela Unifacisa - Centro  
Universitário.

APROVADO EM 07/12/2020

BANCA EXAMINADORA:

  
Prof.<sup>a</sup> da Unifacisa, Lizianne P. M Souto,  
Ms.  
Orientadora

  
Prof.<sup>o</sup> da Unifacisa, Daniel A. C. M de  
Souza, Ms.

  
Prof.<sup>o</sup> da Unifacisa, Jemerson F. Damásio,  
Ms.

Dedico este trabalho à minha família KALUNDA MUTABA e aos meus filhos que são bênçãos vindas de Deus.

## **AGRADECIMENTOS**

A JESUS CRISTO o criador de todas as coisas visíveis e invisíveis, por ter me dado saúde e força para superar as dificuldades.

À Unifacisa, seu corpo docente, direção e administração que oportunizaram a janela que hoje vislumbro um horizonte superior, pelo ambiente criativo e amigável que proporciona a confiança no mérito e ética aqui presente.

À professora Lizianne Souto pelo trabalho de revisão da redação.

Aos meus pais, pelo amor, incentivo e apoio incondicional apesar da longa distância que nós se para.

E a todos que direta ou indiretamente fizeram parte da minha formação, o meu muito obrigado.

*“De tudo o que se tem ouvido, o fim é: Teme a Deus, e guarda os seus mandamentos;  
porque isto é o dever de todo o homem.”*

Rei Salomão (Ecle. 12.13)

## RESUMO

**Introdução** O presente trabalho propõe a investigação de modelos de aprendizagem de máquina aplicados ao problema do diagnóstico de hipertensão arterial. **Objetivo** O trabalho possui como objetivo investigar quais modelos, dentre os selecionados, alcançam melhores taxas de acerto de classificação para o problema. **Metodologia** A metodologia consistiu na execução das seguintes etapas: coleta e pré-processamento de dados, onde o conjunto de dados foi tratado com o objetivo de aumentar a qualidade dos dados, seguido da classificação e, por fim, avaliação dos modelos. **Resultado** Utilizou-se 4 modelos de aprendizado: Redes Bayesianas, Logitboost, Multilayer-Perceptron e J48. Todos os modelos apresentaram acurácia superior a 75%, dentre eles, o modelo de Redes Bayesianas atingiu o melhor resultado; a sua taxa de acerto foi superior a 82%, demonstrando viabilidade em aplicações de previsão de diagnóstico de hipertensão arterial. **Conclusão** A aplicação dos modelos de aprendizado a máquina no diagnóstico de hipertensão arterial apresentou os resultados considerados, podendo auxiliar na detecção dos pacientes com a pressão alta e os que não tem, isso vai ajudar os profissionais de saúde na tomada de decisões.

**PALAVRAS-CHAVE:** Aprendizagem de máquina. Diagnóstico de hipertensão arterial. Redes bayesianas.



## RESUMÉ

**Introduction** Le présent travail propose l'investigation de modèles d'apprentissage automatique appliqués au problème du diagnostic de l'hypertension artérielle. **Objectif** Le travail vise à rechercher quels modèles, parmi les sélectionnés, obtiennent les meilleurs taux d'exactitude de classification pour le problème.

**Méthodologie** La méthodologie consiste à réaliser les étapes suivantes: collecte et prétraitement des données, où l'ensemble de données était traité dans le but d'augmenter la qualité des données, suivi de la classification et, enfin, de l'évaluation des modèles. **Resultat** Quatre modèles d'apprentissage ont été utilisés: les réseaux bayésiens, Logitboost, Multilayer-Perceptron et J48. Tous les modèles ont montré une précision supérieure à 75%, parmi eux, le modèle des réseaux bayésiens a obtenu le meilleur résultat; son taux de réussite était supérieur à 82%, démontrant la faisabilité des applications de prédiction du diagnostic d'hypertension artérielle.

**Conclusion** L'application de modèles d'apprentissage automatique dans le diagnostic de l'hypertension artérielle a montré les résultats considérés, qui peuvent aider à la détection des patients souffrant d'hypertension artérielle et de ceux qui n'en souffrent pas, cela aidera les professionnels de santé dans la prise de décision.

**MOTS-CLÉS:** Apprentissage automatique. Diagnostic d'hypertension artérielle. Réseaux bayésiens.

## LISTA DE ILUSTRAÇÕES

Figura 1 - Equipamento Esfigmomanômetro para Medição .....	24
Figura 2 - Monitorização Ambulatorial da Pressão Arterial .....	25
Figura 3 - Monitorização Residencial de Pressão Arterial .....	26
Figura 4- Redes Neurais Biológicas .....	28
Figura 5 - Arquitetura da Rede neural MLP .....	29
Figura 6 - Exemplo de uma árvore de decisão .....	32
Figura 7 - Weka GUI Chooser .....	34
Figura 8 - Dados carregados no Weka .....	35

**LISTA DE QUADROS**

Quadro 1 - Matriz de Confusão ..... 21

Quadro 2 - Matriz de Confusão ..... 44

## LISTA DE TABELAS

Tabela 1 - Classificação diagnóstica da hipertensão arterial .....	26
Tabela 2 - Cleveland UCI Repository .....	35
Tabela 3 - Definição dos atributos da Cleveland Heart Disease .....	36
Tabela 4 - Acurácia do algoritmo Rede Bayesiana Experimento 1 .....	39
Tabela 5 - Acurácia do algoritmo LogitBoost Experimento 2 .....	40
Tabela 6 - Acurácia do algoritmo Multi-layer Perceptron 3 .....	41
Tabela 7 - Acurácia do algoritmo J48 4 .....	42
Tabela 8 - Resultados de classificação final dos quatros modelos .....	43
Tabela 9 - Comparação dos resultados trabalhos relacionados .....	44
Tabela 10 - Estatística descritiva .....	45
Tabela 11 - Amostragens .....	45

## LISTA DE ABREVIATURAS E SIGLAS

AUC	<i>Area Under The Curve</i> , no inglês, ou Área abaixo da Curva
AD	Árvore de Decisão
AM	Aprendizagem de Máquina
ARFF	<i>Attribute-Relation File Format</i> , ou Arquivo no Formato Atributo Relação.
CSV	<i>Comma Separated Values</i> , ou Arquivo de Valores separados por Vírgula
FP	Falso Positivo
FN	Falso Negativo
HAS	Hipertensão arterial sistêmica
HDL	<i>High Density Lipoprotein</i> , ou Lipoproteína de Alta Densidade
IA	Inteligência Artificial
GPL	<i>General Public License</i>
LDL	<i>Low Density Lipoprotein</i> , ou Lipoproteína de Baixa Densidade
MAPA	Monitorização Ambulatorial da Pressão Arterial
MLP	Multi-layer Perceptron
ML	<i>Machine Learning</i> , no inglês, ou Aprendizagem de Máquina
MRPA	Monitorização Residencial de Pressão Arterial
MS	Ministério da Saúde
KDD	<i>Knowledge-Discovery in Databases</i> , ou Processo de Extração de Informações de Base de Dados
PA	Pressão Arterial
SIM	Sistema de Informações sobre Mortalidade
SGBD	Sistema de Gerenciamento de Banco de Dados
VN	Verdadeiro Negativo
VP	Verdadeiro Positivo

## SUMÁRIO

<b>1</b>	<b>INTRODUÇÃO</b>	<b>15</b>
<b>2</b>	<b>FUNDAMENTAÇÃO TEÓRICA</b>	<b>17</b>
2.1	INTELIGÊNCIA ARTIFICIAL	17
2.1.1	Aprendizado de Máquina	18
2.1.1.1	<i>Aprendizado</i>	19
2.1.1.2	<i>Aprendizado Não-Supervisionado</i>	19
2.1.2	Aprendizado de Máquina na Saúde	20
2.1.3	Medidas de Desempenho	20
2.1.3.1	<i>Matriz de Confusão</i>	21
2.1.3.2	<i>Sensibilidade</i>	21
2.1.3.3	<i>Especificidade</i>	22
2.1.3.4	<i>Acurácia (taxa de Acerto)</i>	22
2.2	DIAGNÓSTICO E CLASSIFICAÇÃO DA HIPERTENSÃO ARTERIAL	22
2.2.1	Medida da Pressão Arterial	23
2.2.1.1	<i>Medida indireta da Pressão Arterial</i>	23
2.2.1.2	<i>Medida Domiciliar e Automedida da Pressão Arterial</i>	24
2.2.2	Classificação de Diagnóstico	26
2.3	MODELOS DE APRENDIZADO DE MÁQUINA	27
2.3.1	Redes Neurais Artificiais (RNAs)	27
2.3.1.1	<i>Multi-layer Perceptron (MLP)</i>	28
2.3.2	LogitBoost (LB)	30
2.3.3	Rede Bayesiana (RB)	30
2.3.4	Árvore de Decisão	31
2.3.5	Algoritmo J48	32
<b>3</b>	<b>METODOLOGIA</b>	<b>33</b>
3.1	CLASSIFICAÇÃO DA PESQUISA	33
3.2	WEKA	33
3.2.1	Exploração de dados Usando Weka	34
3.3	COLETA DE DADOS	35
3.3.1	Pré-processamento	36
3.4	MODELOS DE CLASSIFICAÇÃO	38
<b>4</b>	<b>RESULTADOS E DISCUSSÃO</b>	<b>39</b>
4.1	EXPERIMENTO 1	39
4.2	EXPERIMENTO 2	40
4.3	EXPERIMENTO 3	40
4.4	EXPERIMENTO 4	41
4.5	DISCUSSÃO	42
4.6	COMPARAÇÃO DO RESULTADO	44

<b>5</b>	<b>CONSIDERAÇÕES FINAIS E TRABALHOS FUTUROS .....</b>	<b>46</b>
	<b>REFERÊNCIAS BIBLIOGRÁFICAS .....</b>	<b>47</b>

## 1 INTRODUÇÃO

O Ministério da Saúde Brasileiro (MS, 2019) define a hipertensão arterial ou pressão alta como uma doença crônica determinada pela presença elevada dos níveis tensionais da pressão sanguínea nas artérias. Esta doença é constatada quando os valores da pressão sanguínea máxima e mínima são iguais ou ultrapassam os 140/90 mmHg - pode ser lido no formato 14/9 (i.e., quatorze por nove).

A hipertensão arterial é um dos principais fatores de risco para a ocorrência de acidente vascular cerebral, infarto, aneurisma arterial e insuficiência renal e cardíaca (MS, 2019). Há uma estimativa de que 15 a 20% da população adulta urbana seja acometida por Hipertensão Arterial Sistêmica (HAS).

Atualmente, não existem causas claras para pressão alta. No entanto, existem vários fatores e condições que podem desempenhar um papel importante em seu desenvolvimento. Nesse contexto, pode-se citar: obesidade, falta de atividade física, dieta rica em sal, estresse, idade, histórico familiar, doença renal crônica, entre outros fatores. Uma vez a hipertensão instalada, as consequências e os prejuízos são grandes para a saúde. Por isso, é fundamental o diagnóstico correto bem como o tratamento precoce (SOCIEDADE BRASILEIRA DE CARDIOLOGIA, 2019).

A complexidade do diagnóstico da hipertensão resulta de vários fatores e condições ambientais, que contribuem para o aumento da pressão arterial. Na maioria dos centros hospitalares brasileiros, não existe uma padronização para o diagnóstico. Sempre que possível, a medida da pressão arterial deve ser realizada fora do consultório médico para esclarecer o diagnóstico e afastar a possibilidade do efeito do avental branco (síndrome do jaleco) no processo de verificação. Portanto, a forma e o local onde se mede a pressão arterial é fundamental no estabelecimento do diagnóstico correto da mesma (SOCIEDADE BRASILEIRA DE CARDIOLOGIA, 2019).

Apesar dos avanços científicos e tecnológicos no controle dessa doença, ocorrido nos últimos anos, um dos problemas atuais refere-se ao diagnóstico correto e claro e, em um segundo momento, à prescrição do tratamento adequado. Ultimamente diversos trabalhos aplicaram modelos de Aprendizagem de Máquina (AM) com intuito de melhorar o diagnóstico de hipertensão arterial. Todavia, quando a Inteligência Artificial (IA) é empregada em contextos clínicos mais complexos, ainda há um caminho longo a ser percorrido (AUSTIN et al., 2017).



Diante desse cenário, este trabalho tem como objetivo realizar uma análise comparativa de modelos de Aprendizagem de Máquina aplicados ao diagnóstico da hipertensão arterial. Para fins de experimentação, será utilizado um *dataset* (i.e., conjunto de dados estruturados) disponível na internet, obtido numa plataforma de hospedagem para projetos e competições de Ciência de Dados. Adicionalmente, pretende-se realizar a análise comparativa dos modelos selecionados e identificar o modelo mais preciso e com mais acertos no diagnóstico. Para isso, serão avaliados dentro de cada modelo os principais hiperparâmetros (i.e., aqueles parâmetros de configuração que mais contribuem para o bom desempenho do modelo).

Pretende-se que o resultado desse trabalho tenha contribuições significativas na área de computação e medicina. No primeiro caso, analisando quatro modelos de Aprendizagem de Máquina, avaliando-os dentro de um contexto específico os aspectos positivos e negativos do emprego de cada um. No segundo caso, fornecendo subsídios úteis para automatizar o diagnóstico da hipertensão arterial, auxiliando profissionais de saúde na diminuição da incerteza na tomada de decisão, reduzindo o número de casos diagnosticados como falsos negativos, por exemplo.

De acordo com o contexto, objetivo e questões de pesquisa do presente trabalho, este encontra-se organizado da seguinte forma: no Capítulo 2, são apresentados os principais conceitos que facilitam o entendimento da presente pesquisa; no Capítulo 3, encontra-se a metodologia científica deste estudo. No Capítulo 4 são apresentados os resultados e discussão associados à aplicação dos modelos de Aprendizagem de Máquina no diagnóstico de hipertensão arterial. Capítulo 5, são apresentadas as principais conclusões deste trabalho, bem como sugestões de trabalhos futuros.

## 2 FUNDAMENTAÇÃO TEÓRICA

Neste capítulo, são apresentados os principais conceitos que possibilitam o entendimento da presente pesquisa. São descritos os conceitos tecnológicos envolvendo a área de Inteligência Artificial na seção 2.1, enquanto na seção 2.2, são mostrados, de forma resumida, os problemas relacionados ao diagnóstico da hipertensão arterial e suas classificações. Finalmente, na seção 2.3, são descritos os modelos de Aprendizagem de Máquina considerados neste trabalho.

### 2.1 INTELIGÊNCIA ARTIFICIAL

Por definição, a Inteligência Artificial (IA) é uma área da ciência da computação que teve seu início nos anos 1956 com objetivo de trazer a forma de aprendizagem natural para os computadores. Ela contribuiu muito para o avanço da tecnologia, hoje em dia, várias organizações ou empresas têm a facilidade de detectar e resolver problemas, tomando boas decisões tudo isso no espaço de pouco tempo (RUSSEL; NORVIG, 2004).

Na literatura é possível encontrar várias definições para Inteligência Artificial, algumas são consideradas como subjetivas e outras mais práticas, em conformidade com a área de atuação.

Encontram-se listadas abaixo algumas áreas de pesquisa da IA (RUSSEL; NORVIG, 2004):

- a) Solucionadores de Problemas: uma abordagem examinadora que, os algoritmos aprendem a resolver problemas de maneira semelhante ao ser humano com tempos reduzidos e, indicam um caminho com boa chance de resultado;
- b) Raciocínio Lógico: um processo de estruturação do pensamento de acordo com as normas da lógica que permite chegar a uma determinada conclusão ou resolver um problema;
- c) Processamento de Linguagem Natural: a interpretação e compreensão de textos;
- d) Robótica e Visão: inclui manipulação de objetos, sequenciamento de tarefas, reconhecimento e detecção de padrões;
- e) Programação Automática: consiste em gerar código de programação de computadores de maneira automática;

- f) Aprendizagem: a forma que o sistema aprende com sua própria experiência, desenvolvendo seu desempenho. Nesta área é que se baseia este trabalho;
- g) Sistemas Especialistas: armazenam o conhecimento de uma área específica de atuação, utilizando-o como suporte à tomada de decisão.

### 2.1.1 Aprendizagem de Máquina

Atualmente, vive-se em uma época considerada como “Era da informação” ou “Era digital”, isso porque dados são produzidos todos os dias em uma quantidade massiva, vindos de: transações comerciais, resultados de pesquisas científicas, tráfego na internet, armazenamento de informações como conversas na rede. A quantidade de dados gerados só tende a aumentar, e com o progresso da tecnologia, estes dados podem ser armazenados em grandes repositórios de dados estruturados como *Armazém de Dados*, Bancos de Dados, estruturas de *blockchain*, entre outros meios. Mesmo assim, esta imensa quantidade de dados não possui importância significativa se estes dados não puderem ser analisados e entendidos devidamente.

Nesse contexto, surge uma área intitulada de Aprendizagem de Máquina (em inglês *Machine Learning* - *ML*), pelo qual métodos são aplicados para auxiliar os computadores a aprender a resolver problemas com a experiência passada. Para esse fim, aplica-se um princípio de inferência denominado indução. Onde os modelos de AM aprendem a induzir funções ou hipóteses capazes de solucionar determinados problemas (FACELI et al., 2011).

Duas categorias são observadas em AM, a primeira é a preditiva e a segunda descritiva. A preditiva tem como intuito encontrar função, baseada nos dados treinados, que por sua vez, quando for usada tenha possibilidade de prever um rótulo ou valor que define um novo exemplo, com base nos valores de seus atributos de entrada. Esse padrão é chamado de aprendizagem supervisionada, onde há a presença de um “supervisor externo”, que tem conhecimento da saída esperada. O supervisor avalia o desempenho do modelo (FACELI et al., 2011).

Por sua vez, na descritiva, os dados são categorizados conforme as suas características. O padrão seguido por modelos chama-se de aprendizado não supervisionado, útil em atividades de mineração de dados para análises futuras (FACELI et al., 2011).

### **2.1.1.1 Aprendizado**

A aprendizagem a máquina auxilia o ser humano para resolver problemas e a maioria deles são tratados de forma supervisionada, e observa-se que o aprendizado supervisionado tem uma característica básica, isto é, os dados utilizados para o treinamento incluem a resposta desejada chamada de classes a serem previstas.

A abordagem de Aprendizado Supervisionado consiste em utilizar uma série de exemplos (chamados de instâncias), já marcados, para induzir um modelo que seja capaz de classificar novas instâncias de forma precisa, com base no aprendizado obtido com os dados de treinamento. É comum que este modelo seja chamado também de classificador. Portanto, é importante que exista um conjunto de dados de treinamento de qualidade, para que o modelo criado possa ser capaz de prever novas instâncias de forma eficiente (BRINK; RICHARDS, 2013).

### **2.1.1.2 Aprendizado Não-Supervisionado**

Classificado como uma tarefa descritiva, Aprendizado Não-Supervisionado, uma abordagem que é utilizada para identificar informações consideradas nos dados não-categorizados. Essa abordagem não requer a presença de um supervisor, não há uma avaliação dos resultados (BRINK; RICHARDS, 2013).

Os modelos utilizados no aprendizado Não-Supervisionado não precisam de referência ou alguns critérios determinados para seguir, isto significa que não existe classificação para o conjunto de dados, o resultado obtido pelo modelo na saída não é esperado. Os algoritmos processam os dados com base em características semelhantes, os dados são agrupados no mesmo grupo quando apresentam as mesmas características e os distintos vão ficar em grupos diferentes (FURNKRANZ; GAMBERGER; LAVRAC, 2012).

Este trabalho tem como foco o Aprendizado Supervisionado, especialmente os modelos de classificação, a abordagem Não-Supervisionado não será detalhada nesta pesquisa.

### 2.1.2 Aprendizagem de Máquina na Saúde

Diversos estudos têm sido desenvolvidos no setor de aprendizagem de máquina aplicados à saúde. A evolução das pesquisas nessa área ocorre devido à necessidade de ferramentas que possam apoiar o diagnóstico médico de forma eficaz e eficiente.

O estudo de Shubhankar (2019), por exemplo, teve como objetivo determinar os fatores de risco para contrair doenças cardíacas, utilizando aprendizagem de máquina para construir modelos preditivos para a ocorrência de doenças a partir de uma base de dados disponível na internet. O modelo desenvolvido fez uso dos dados da pesquisa *Cleveland Heart Disease UCI Repository*, na qual foram coletados dados pessoais de cerca de 303 indivíduos. Os resultados obtidos indicaram que o uso de dados de saúde para a construção de modelos preditivos apresentou o melhor resultado dentre os testes realizados, com uma acurácia de predição de 80.32% com o modelo Regressão Logística. Dentre os dados de saúde, a atividade física e a presença de algumas condições de saúde foram fortes preceptores individuais.

No contexto brasileiro, Oliveira et al. (2017) desenvolveram modelos preditivos de diabetes não diagnosticada, a partir de dados de 12.447 adultos entrevistados para o Estudo Longitudinal de Saúde do Adulto (ELSA). Os modelos que atingiram os melhores resultados foram Redes Neurais Artificiais 75.24% e Regressão Logística 74.98%. A frequência de diabetes não diagnosticada foi de 11%. Entre os 403 indivíduos do conjunto de dados de teste que tinham diabetes não diagnosticada, 274 foram identificados como casos positivos.

### 2.1.3 Medidas de Desempenho

A avaliação de desempenho dos classificadores se realiza através dos indicadores de desempenho. Nesta seção são apresentadas algumas medidas de desempenho mais comumente utilizadas no contexto de doenças, como a hipertensão arterial e, em algoritmos de classificação.

### 2.1.3.1 Matriz de Confusão

O Hossin, (2015) define a matriz de confusão como uma tabela que proporciona a percepção da performance de um modelo de classificação, carregando os valores previstos em linhas e colunas. As linhas da tabela correspondem às instâncias de uma classe esperada e as colunas representam as instâncias da classe atual.

No Quadro 1 é mostrada a matriz de confusão com as classes desejadas e as classes previstas. Lembrando que a classe positiva corresponde à presença da doença e a classe negativa, a sua ausência.

Quadro 1 - Matriz de Confusão

	Positivos previstos	Negativos previstos
Positivos originais	VP	FP
Negativos originais	FN	VN

Fonte: (HOSSIN 2015)

Alguns valores representativos da matriz de confusão:

- a) Verdadeiro Positivo (VP): no contexto de hipertensão arterial, VP representa a taxa de exemplos na base de dados classificados corretamente como apresentando a doença;
- b) Falso Positivo (FP): corresponde aos exemplos na base de dados que foram classificados incorretamente como apresentando a doença;
- c) Verdadeiro Negativo (VN): corresponde aos exemplos no banco de dados que não possuem a doença e foram classificados corretamente como negativo;
- d) Falso Negativo (FN): corresponde aos exemplos na base de dados que apresentaram a doença e foram classificados incorretamente como negativos.

### 2.1.3.2 Sensibilidade

De acordo com Hossin et al., (2015) a Sensibilidade é a taxa de Verdadeiros Positivos (VP), ela representa a quantidade de acerto na classe positiva, isto é, a parte de casos positivos classificados de maneira correta. A equação aplicada para esta medida é a seguinte:

$$\text{Sensibilidade} = VP / (VP + FN)$$

### **2.1.3.3 Especificidade**

Especificidade, por definição, é a taxa de verdadeiros Negativos (VN), ela corresponde à taxa de acerto na classe negativa. A Especificidade avalia a proporção de casos negativos que foram classificados corretamente (Hossin et al., 2015). A equação aplicada para esta medida é a seguinte:

$$\text{Especificidade} = VN / (VN + FP)$$

### **2.1.3.4 Acurácia (Taxa de acerto)**

Nos problemas ligados a classificação, a acurácia corresponde à percentagem de instâncias que o algoritmo previu corretamente de modo geral, em outros termos, a taxa de exemplos positivos e negativos corretamente classificados (Hossin et al., 2015). A equação responsável para esta medida é a seguinte:

$$\text{Acurácia} = (VP + VN) / (VN + FP + VP + FN)$$

## **2.2 DIAGNÓSTICO E CLASSIFICAÇÃO DA HIPERTENSÃO ARTERIAL**

A realização de diagnóstico da Hipertensão Arterial é feita em qualquer unidade de saúde ou até domiciliar, os médicos e profissionais de saúde têm o dever de fazer a medição e avaliação de pressão arterial (PA) seguindo as normas estabelecidas.

Segundo o Ministério de Saúde (MS, 2019), o diagnóstico de hipertensão arterial baseia-se num procedimento acessível, a medição da pressão arterial, que é de uma grande responsabilidade de classificar um paciente como hipertenso ou não. Um diagnóstico cujo resultado errado, tem consequências catastróficas. Um paciente que é hipertenso, e que é diagnosticado como não hipertenso, é privado do tratamento, conseqüentemente, o paciente hipertenso, classificado como não hipertenso, será submetido a tratamentos inúteis e até perigosos à saúde.

Assim, com a realização do diagnóstico prévio, um tratamento adequado e contínuo são armas eficazes no combate da hipertensão arterial.

### **2.2.1 Medida da Pressão Arterial**

A Sociedade de Cardiologia Brasileira (SCB, 2019), recomenda-se um lugar calmo para efetuar a medição da Pressão Arterial, principalmente nas unidades de saúde e os profissionais da área de saúde são responsáveis por esse procedimento.

Para um bom diagnóstico, devem ser considerados os diferentes tipos de medição, apresentados nas subseções a seguir.

#### **2.2.1.1 Medida Indireta da Pressão Arterial**

Com o Esfigmomanômetro que é o instrumento ideal para medida indireta da PA, o processo é realizado em todas as unidades hospitalares (Figura 1). O aparelho deve ser testado, calibrado de maneira regular antes e depois de uso (NOGUEIRA et al., 1998). A seguir encontram-se alguns passos elementares da medida indireta da pressão arterial para detecção, avaliação e controle de hipertensos (SOCIEDADE BRASILEIRA DE CARDIOLOGIA, 2019):

- a) O paciente precisa ter uma explicação clara do procedimento;
- b) O avaliador(a) precisa afirmar que o paciente:
  - Não está com a bexiga cheia;
  - Não realizou exercícios físicos no intervalo de 60-90 minutos antes da avaliação;
  - Não consumiu alimentos inadequados tais como: alimentos salgados, líquidos alcoólicos, café, ou fumou até 30 minutos anteriormente da medida;
- c) Deixar o paciente descansar por 10 minutos em ambiente calmo, com temperatura agradável;
- d) Localizar a artéria braquial por palpação;
- e) Colocar o manguito firmemente cerca de dois a três centímetros acima da antecubital;
- f) Fixar os olhos no mesmo nível da coluna de mostrador manômetro;
- g) Tocar o pulso radial, encher o manguito até a desapareção do pulso com intuito de relevar o nível da pressão sistólica, depois, desinflar o manguito de maneira rápida e esperar mais ou menos um minuto antes de encher novamente etc.



A Figura 1 mostra o equipamento Esfigmomanômetro que serve na medição de pressão arterial.

Figura 1 - Equipamento Esfigmomanômetro para medição



Fonte: Dental Access (2020)

### **2.2.1.2 Medida Domiciliar e Automedida da Pressão Arterial**

Esse tipo de medida é realizado fora do consultório, considerada como uma medida eficaz em que o paciente coopera de forma ativa no processo e os diferentes valores registrados são apresentados para o médico e este vai calcular a média para determinar o exato valor da pressão arterial.

Dois métodos são recomendados para medir a pressão arterial, primeiro a Monitorização Ambulatorial da Pressão Arterial (MAPA), que consiste no uso do equipamento no braço do paciente durante 24 horas para monitorar a pressão, a cada 15-30 minutos ocorre uma medição. À noite, esse intervalo aumenta para 20-30 minutos. De acordo com o II Consenso de MAPA, o método apresenta melhor correlação com risco cardiovascular do que a medida da pressão arterial de consultório (MS, 2012). A Figura 2 mostra o equipamento de Monitorização

Ambulatorial da Pressão Arterial (MAPA), feita geralmente a cada 20 minutos durante o período de 24 horas.

Figura 2 - (MAPA) Monitorização Ambulatorial da Pressão Arterial



Fonte: Instituto de Cardiologia do Lago (2020)

O segundo método é a Monitorização Residencial de Pressão Arterial (MRPA). Onde é constatado a colaboração do paciente no processo, ele mede três vezes pela manhã e três vezes à noite, durante cinco dias. O paciente segue as orientações para que o uso do aparelho seja feito de forma adequada. Assim como o método MAPA, depois desse tempo os dados registrados serão apresentados para o médico e a média é calculada para a determinação da PA fora do consultório. (SOCIEDADE BRASILEIRA DE CARDIOLOGIA, 2019).

Esses procedimentos são úteis para (SOCIEDADE BRASILEIRA DE CARDIOLOGIA, 2019):

- a) Identificar a hipertensão do avental branco (hipertensão de consultório isolada);
- b) Avaliar a eficácia da terapêutica anti-hipertensiva;
- c) Estimular a adesão ao tratamento; e reduzir custos.

A Figura 3 mostra o equipamento MRPA que serve na medição de pressão arterial.

Figura 3 - (MRPA) Monitorização Residencial de Pressão Arterial



Fonte: Instituto de Cardiologia do Lago (2020)

### 2.2.2 Classificação de Diagnóstico

Os pacientes diagnosticados são classificados em diferentes categorias com base no valor da pressão arterial obtido durante a vigília pela MAPA ou MRPA.

O número de pressão analisado é opcional, a classificação deste depende da sua medição. Considera-se normal o paciente adulto cuja idade seja maior a 18 anos, números inferiores a 85 mmHg de pressão diastólica e inferiores a 130 mmHg de pressão sistólica. Já que, o paciente normal limítrofe de 130 - 139 mmHg/85 - 89, precisa de ser beneficiado com as medidas preventivas. Na Tabela 1 é apresentada a classificação diagnóstica da hipertensão arterial (SOCIEDADE BRASILEIRA DE CARDIOLOGIA, 2019).

**Tabela 1 - Classificação diagnóstica da hipertensão arterial (> 18 anos de idade) (Continua)**

PAD (mmHg)	PAS (mmHg)	CLASSIFICAÇÃO
<85	<130	Normal
85-89	130-139	Normal limítrofe
90-99	140-159	Hipertensão leve (estágio 1)
100-109	160-179	Hipertensão moderada (estágio 2)

**Tabela 1 - Classificação diagnóstica da hipertensão arterial (> 18 anos de idade) (Conclusão)**

PAD (mmHg)	PAS (mmHg)	CLASSIFICAÇÃO
≥110	≥180	Hipertensão grave (estágio 3)
<90	≥140	Hipertensão sistólica isolada

Fonte: Sociedade Brasileira de Cardiologia (2019)

Os valores apresentados na Tabela 1 possibilitam a classificação dos pacientes em categorias diferentes, normal, normal limítrofe, hipertensão leve (estágio 1), hipertensão moderada (estágio 2), hipertensão grave (estágio 3) e, por fim, a hipertensão sistólica isolada. Define-se efeito normal com o valor obtido abaixo do percentil 90, normotensão; quando os valores atingem os percentis 90 e 95 normal limítrofe; e por fim, os valores acima do 95 o paciente é considerado hipertenso, assim ele necessita de atenção de investigação de causas secundárias e tratamento apropriado, já que se encontra na zona de risco (SOCIEDADE BRASILEIRA DE CARDIOLOGIA, 2019).

## 2.3 MODELOS DE APRENDIZAGEM DE MÁQUINA

Diversos modelos são encontrados na literatura para aplicação em AM, entre os quais destaca-se: Redes Neurais Artificiais, *Multi-layer Perceptron*, *LogitBoost* (LB), Redes Bayesianas (BN) e Árvore de Decisão (AD).

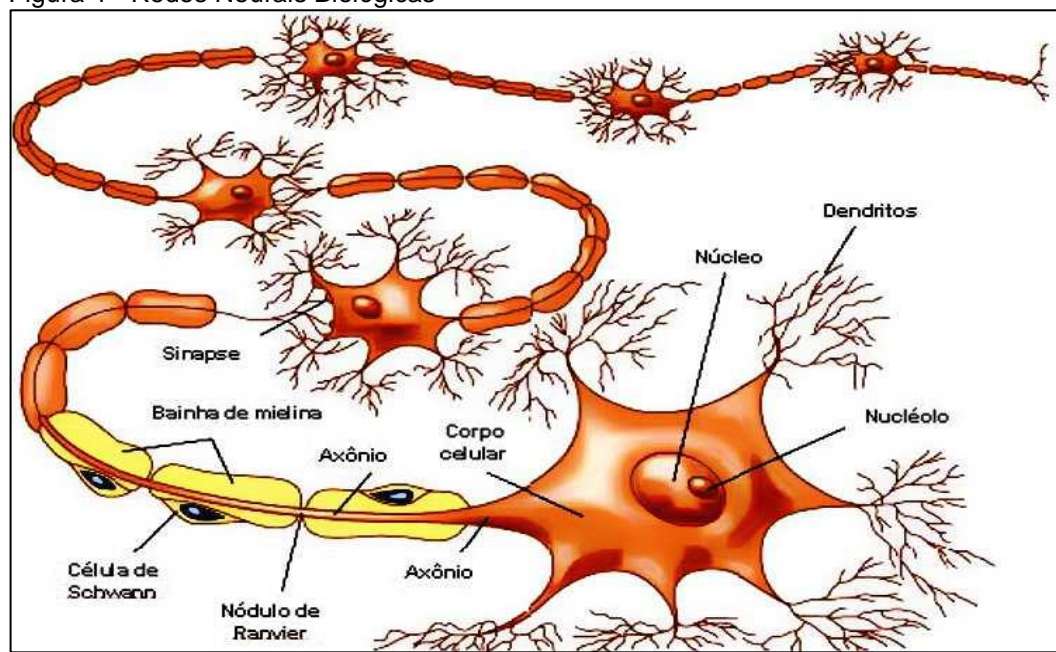
### 2.3.1 Redes Neurais Artificiais (RNAs)

Por definição, as Redes Neurais Artificiais são modelos de aprendizagem de máquina desenvolvidos com a inspiração do cérebro humano, elas possuem capacidade de aprender e gerenciar processos de diferentes áreas (HYAKIN, 2001).

Fala-se de RNA quando um conjunto dos neurônios são conectados entre si, e as ligações entre neurônios são influenciadas pelos pesos que possuem a informação dentro da RNA. Através desta relação, a saída de um neurônio consiste na entrada de outro, como demonstrado na Figura 4, a qual é um exemplo de uma rede neural biológica (CERA, 2005). Essa é a estrutura de uma rede neural biológica, composta por três partes fundamentais: os dendritos, tens como objetivo receber as informações

vindo de outros neurônios, o corpo celular, assume a responsabilidade de tratar essas informações, e por fim axônio, ele tem como função compartilhar as informações processadas por outros neurônios.

Figura 4 - Redes Neurais Biológicas



Fonte: MARCOS (2014)

Na estrutura da rede neural artificial, um neurônio artificial é preparado para processar uma única tarefa, nas entradas são recebidos apenas um tipo de sinal ou informação. Uma rede neural é composta de várias conexões entre neurônios, e cada um dele possui várias entradas que facilita a percepção de diferentes sinais vindo de outros neurônios, esta é a característica da rede neural, processar mais informações e fornecer mais resultados devido a ligação de várias células em rede (CERA, 2005).

### 2.3.1.1 Multi-layer Perceptron (MLP)

As redes MLP são caracterizadas pelas elevadas possibilidades de aplicações em diversos tipos de problemas, relacionados com as mais diferentes áreas de conhecimento, consideradas também como uma arquitetura multifuncional quanto à sua aplicabilidade. Dentre essas potenciais áreas, têm-se os seguintes destaques (FACELI, 2011):

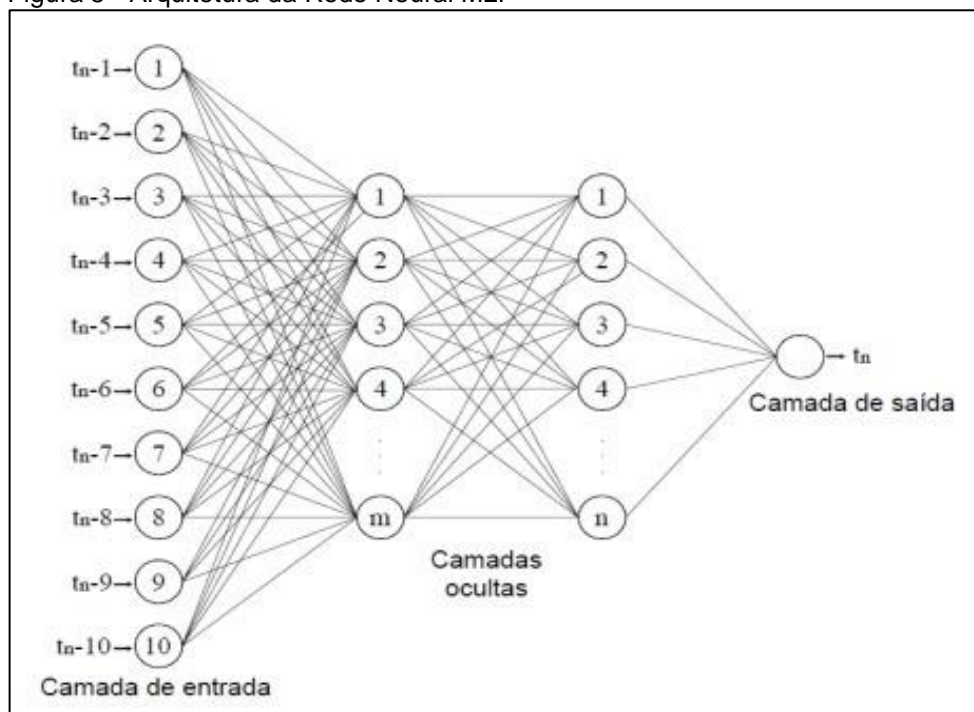
- a) Reconhecimento de padrões;
- b) Identificação e controle de processos;

- c) Previsão de séries temporais;
- d) Otimização de sistemas etc.

Nunes et al., (2016), afirmam que as redes MLP pertencem à arquitetura *feedforward* que é formada por várias camadas múltiplas e algumas delas são escondidas, e os treinamentos são realizados de forma supervisionada.

Na Figura 5 é possível observar a arquitetura de uma Rede Neural. Formada por um ou mais camadas de neurônios. As camadas estão constituídas por neurônios alinhados e ligados entre si. A entrada de uma camada intermediária “t” é a saída da camada anterior, isto é, “t-1”, na saída a mesma camada será a camada da entrada seguinte “t+1” que será a camada de saída caso seja a última na linha de conexão. O fluxo de informações na estrutura da rede é inicializado na camada de entrada, percorre em seguida as camadas intermediárias, e finalizado na camada neural de saída.

Figura 5 - Arquitetura da Rede Neural MLP



Fonte: MARCOS (2014)

A partir da ilustração, verifica-se que a rede recebe as  $n$  entradas  $\{t_1, t_2, t_3, t_4, t_5, t(n)\}$ , representando camadas de entradas,  $\{1, 2, 3, 4...m\}$  e  $\{1, 2, 3, 4...n\}$ , representando camadas ocultas e  $\{t_n\}$  representando camada de saída. Da entrada à saída. O processo prediz como resposta o respectivo valor esperado para o padrão



detectado fornecido pelo seu neurônio de saída ( $t_n$ ). Assim, durante o processo de treinamento, a rede tentará ajustar as suas matrizes de pesos visando minimizar o erro produzido durante o processo.

### 2.3.2 LogitBoost (LB)

O modelo *LogitBoost* é um dos algoritmos de *boosting* cujo objetivo é de ajustar modelos suplementares mediante a uma melhoria de um critério, ou funcional de custo. *LogitBoost* é um algoritmo de impulso formulado por Jerome Friedman, Trevor Hastie e Robert Tibshirani (ano).

De acordo com Friedman et al., (2000), o *LogitBoost* posiciona menor ênfase nas amostras que são classificadas de uma forma incorreta. Quando ele é aplicado na prática, apresenta-se um melhor desempenho em amostras “ruídos”. Os ruídos são interferências entre dados coletados em locais não controlados, que afetam de forma significativa a performance dos modelos de classificação em termos de acurácia, no tempo determinado para o treinamento e na complexidade de resolução dos problemas.

Nesse contexto, considera-se ruídos, os erros introduzidos aos valores dos atributos, entre outros, valores incorretos, faltantes ou desconhecidos e também os ruídos de classes que são instâncias rotuladas incorretamente, isso ocorre quando a instância é duplicada na base com as classificações distintas, e por fim, outro erro de classificações quando a instância é avaliada com uma classe que não pertence.

O método *boosting*, possui vários algoritmos, dentre eles encontram-se: Algoritmo *AdaBoost Real*, Algoritmo *AdaBoost* Discreto, e por último Algoritmo *LogitBoost*.

### 2.3.3 Rede Bayesiana (RB)

Conhecidas como redes causais, as redes bayesianas representam as relações de probabilidade condicional que se refere a uma probabilidade de determinado evento *A* que ocorre com a influência de outro evento *B* representada por  $P(A/B)$ , isto é, “probabilidade de evento *A* ocorrer com a dependência de evento *B*”. Surgiu em 1980 com objetivo de facilitar o trabalho de predição na área de aprendizagem de máquina.

Segundo Russel et al., (2004), as Redes Bayesianas são modelos de representação probabilísticos que são usados para caracterizar relações existentes entre variáveis. Por outro lado, as Redes Bayesianas estabelecem um padrão gráfico que apresenta de maneira compreensível as relações de causalidade das variáveis de um sistema, isso é considerado como uma das vantagens das RBs.

Encontra-se o interesse no uso das redes probabilísticas para o tratamento de incerteza ligado ao diagnóstico de hipertensão arterial. Suas vantagens se apresentam da seguinte maneira: facilitar a representação e manipulação da incerteza com conceitos matemáticos; As Redes Bayesianas permitem a realização possíveis de indução probabilística, diagnóstico, intercausal ou misto etc.

### **2.3.4 Árvore de Decisão**

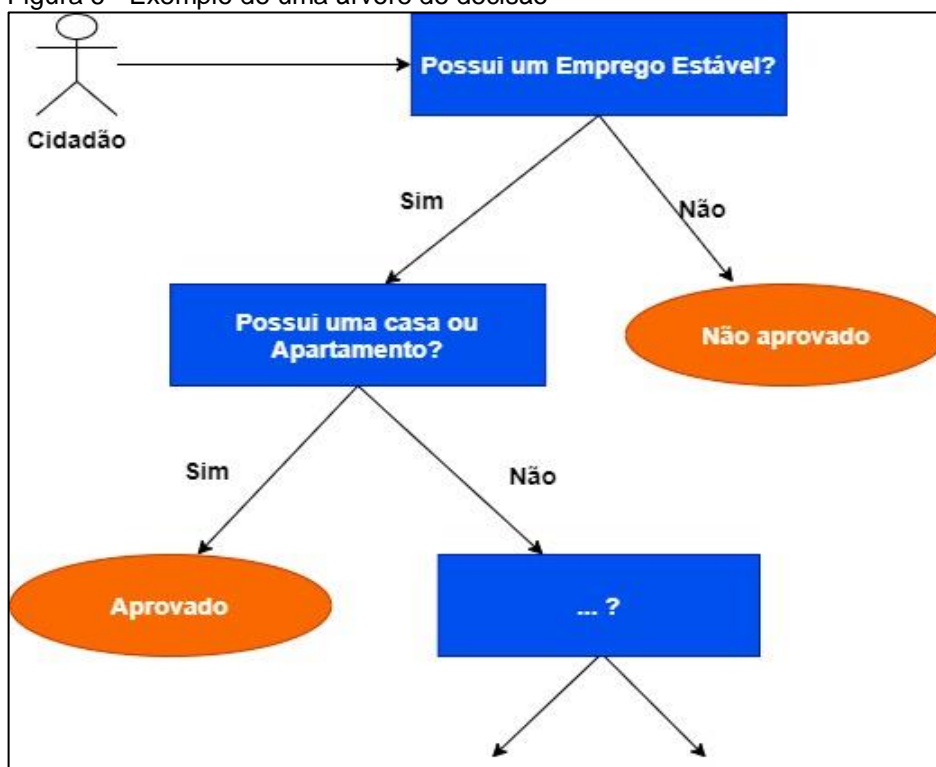
O uso da árvore de decisão tem um número considerável na área de Inteligência Artificial (IA), com a sua técnica de “dividir para conquistar” que tem como objetivo decompor um problema em problemas menores facilitando a tomada de decisão para resolvê-lo.

A estrutura da árvore de decisão se apresenta de maneira seguinte: no topo da árvore encontra-se o nó raiz, seguido dos demais nós que representam os atributos; arcos (ramos), os receptores dos valores dos atributos (um ramo representa um valor do atributo), folhas de árvores chamados de (nodos folha), descrevem as diferentes classes de um conjunto de treinamento, isto é, associação de folha com a classe.

Na Figura 6, é mostrada uma árvore de decisão cujo objetivo é a validação de uma solicitação de conta bancária. A instância 1 verifica se o candidato possui um emprego estável. Caso não possua, a abertura da conta não será aprovada. No caso afirmativo, ele será submetido a outro teste. A instância 2 consiste em verificar se o candidato possui uma casa ou apartamento. Caso seja válido ele recebe uma conta, no caso contrário, o candidato passará por outro teste até alcançar o nível mais baixo.



Figura 6 - Exemplo de uma árvore de decisão



Fonte: O Autor (2020)

### 2.3.5 Algoritmo J48

Considerado como o mais usado da ferramenta Weka, o algoritmo J48 é uma implementação do algoritmo C4.5 que tem a capacidade de tratar os atributos contínuos que são números reais de valores por exemplo: a altura, temperatura ou peso e atributos discretos que representam um conjunto de valores finito ou infinito (QUINLAN, 1993).

Desenvolvido por Ross Quinlan (1993), o J48, uma extensão do algoritmo de classificação anterior de ID3, ele foi implementado em Java e encontra-se disponível na ferramenta Weka. No seu funcionamento o algoritmo J48 cria árvores de decisão com base de dados de treinamento aplicando o conceito de *Entropia*, isto significa, que o atributo mais eficaz para a diminuição da entropia é selecionado e considerado como *raiz* (nó) da árvore, a sub-árvore do nó raiz é formada para cada valor provável deste atributo. Esse processo é repetido considerando a mesma lógica até que a árvore seja formada por completo.

### 3 METODOLOGIA

Neste capítulo são apresentados as técnicas e métodos de investigação que colaboraram para a realização deste trabalho considerando a classificação da pesquisa quanto à natureza, objetivos, e procedimentos técnicos (APPOLINÁRIO; GIL, 2004). Assim como, a descrição da ferramenta usada nos experimentos e o banco de dados utilizado.

#### 3.1 CLASSIFICAÇÃO DA PESQUISA

Em conformidade à natureza, este trabalho classifica-se como uma pesquisa aplicada, uma vez que são aplicados modelos de aprendizagem de máquina a um conjunto de dados para a previsão de diagnóstico médico. Esse conhecimento é aplicado numa realidade bem particular, o diagnóstico de hipertensão arterial.

Em relação aos objetivos, o presente trabalho é classificado como descritivo. A pesquisa visa melhorar o atendimento dos pacientes na realização de diagnóstico de hipertensão arterial sem erro no diagnóstico, com exames mais precisos, prevenir hipertensão arterial, e também garantir a acurácia dos valores obtidos.

Quanto aos procedimentos técnicos, trata-se de um estudo experimental que envolve aplicação de algoritmos classificadores de aprendizagem de máquina para construção de modelos preditivos da hipertensão arterial, utilizando o software *Weka* e dados hospitalares (*dataset*) disponíveis na internet de forma gratuita.

#### 3.2 WEKA

Para a exploração dos dados, não é suficiente ter apenas domínio do conhecimento sobre as variáveis, atributos e outros que são significativos. Na verdade, é preciso ter um instrumento que possa auxiliar no gerenciamento destes. Existem na literatura ferramentas para realização de mineração de dados como, por exemplo, *IBM Intelligent Miner*, *DBMiner*, *MinerSet* e *Weka*. Optou-se pela utilização da ferramenta *Weka* devido sua simplicidade e interface gráfica, que proporcionam facilidade de uso (Figura 7).

A *Weka* (*Waikato Environment for Knowledge Analysis*), desenvolvido na linguagem Java, possui uma coleção de algoritmos de aprendizagem de máquina, de

código aberto. Em 1997, a ferramenta foi implementada pela primeira vez em sua forma moderna. A Weka possui uma GUI (*Graphical User Interface*) que possibilita a interação entre usuário com arquivos de dados e fornecendo resultados visuais. (HALL et al., 2009).

Figura 7 - Weka GUI Chooser

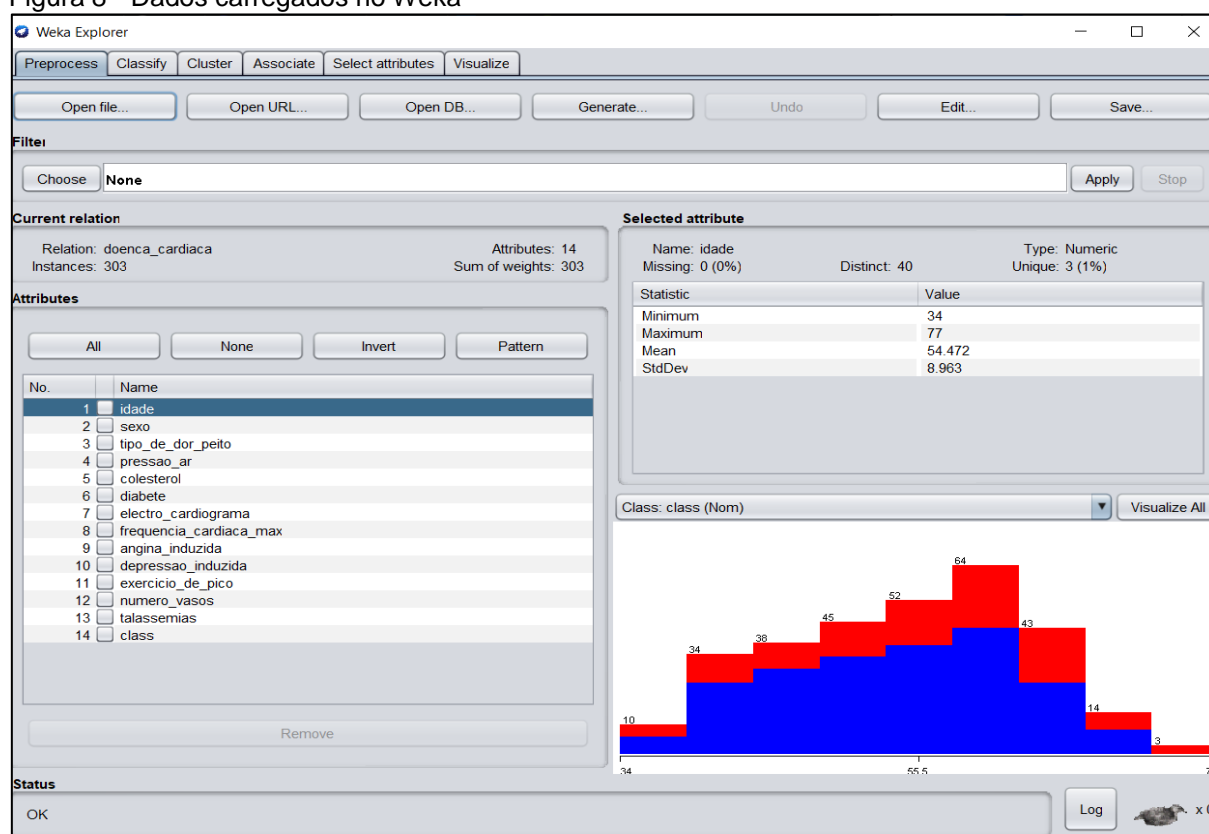


Fonte: O Autor (2020)

### 3.2.1 Exploração de dados Usando Weka

Algumas funcionalidades que a ferramenta contém são apresentadas na interface gráfica, tais como: pré-processamento; classificação, clusterização, associação, e visualização dos resultados (formas gráficas, estatísticas). O Weka utiliza dados no formato ARFF e CSV. A Figura 8 apresenta uma tela, onde arquivo de dados é carregado, e o usuário pode visualizar as estatísticas do conjunto selecionado. A interface oferece a possibilidade de selecionar o algoritmo a ser aplicado e fazer as alterações da configuração deste se necessário

Figura 8 - Dados carregados no Weka



Fonte: O Autor (2020)

### 3.3 COLETA DE DADOS

Nesta seção apresenta-se os dados utilizados nesta pesquisa. Optou-se pela coleta de dados no repositório de *Machine Learning UCI (Machine Learning Repository) Cleveland Clinic*, disponíveis no formato csv, para realização de testes e consolidação dos resultados. A base possui dados relacionados à doença do coração, porém, pode ser utilizada no contexto hipertensão devido a existência de um atributo rótulo da classe que informa se o paciente possui ou não hipertensão.

A Tabela 2 apresenta de maneira resumida o conjunto de dados explorados nesta pesquisa.

**Tabela 2 - Cleveland Heart Disease UCI Repository (Continua)**

Idade	Sexo	Tipo de dor peito	PA	Colesterol	Açúcar no sangue	Electro cardiograma	Frequência cardíaca máx
63	1	3	145	233	1	0	150
37	1	2	130	250	0	1	187

**Tabela 2 - Cleveland Heart Disease UCI Repository (Conclusão)**

Idade	Sexo	Tipo de dor peito	PA	Colesterol	Açúcar no sangue	Electro cardiograma	Frequência cardíaca máx
41	0	1	130	204	0	0	172
56	1	1	120	236	0	1	178
57	0	0	120	354	0	1	163

Fonte: Cleveland Heart Disease (2020)

### 3.3.1 Pré-processamento

Na base de dados original havia 76 atributos, mas para a realização dos experimentos alguns deles foram desconsiderados devido à baixa ou à ausência total de utilidade para o processo de diagnóstico. Isto resultou na seleção de um subconjunto de 14 atributos mais relevantes para a realização dos experimentos.

Em relação aos dados faltantes, percebeu-se 3 campos brancos e foram preenchidos com os valores que mais ocorrem nas colunas respectivas, usou-se a técnica da “*moda*”. Após a realização do pré-processamento o *dataset* manteve informações no total de 165 pacientes considerados positivos e 138 pacientes negativos. O número dos pacientes é de 207 mulheres e 96 homens, com faixa etária de 29 e 77 anos. Na Tabela 3 tem-se uma amostra dos seus registros e colunas. No total, a base conta com 14 colunas e 303 registros de pacientes.

Na Tabela 3, encontram-se os atributos considerados: idade, sexo, tipo de dor no peito, pressão sanguínea em repouso, colesterol, presença de açúcar no sangue (diabete), electro cardio repouso, frequência cardíaca máximo, angina Induzida por exercício, depressão induzida, inclinação de segmento (exercício pico), número de vasos principais, resultado de estresse e, por último o atributo relevante para esta pesquisa, o estado de hipertensão.

**Tabela 3 - Definição dos atributos da Cleveland Heart Disease (Continua)**

Num	Atributo	Descrição	Valores
01	Idade	Indica a idade do paciente no momento do atendimento.	de 29 a 77 anos
02	Sexo	Indica o sexo do paciente	0 - Masculino 1 - Feminino

**Tabela 3 - Definição dos atributos da Cleveland Heart Disease (Conclusão)**

Num	Atributo	Descrição	Valores
03	Tipo de dor no peito	Indica o nível de dor que o paciente sente no peito.	0 - Angina típica 1 - Angina atípica 2 - Dor não anginosa 3 - Assintomático
04	Pressão Arterial	Indica a medida da pressão exercida pelo sangue na parede das artérias, expressando a condição do sistema circulatório do paciente	< 120/80 - Ótima < 130/85 - Normal 130-139/85-89 limítrofe 140-159/90-99 Hipertensão 1 160-179/100-109 Hipertensão 2 >180/>110 Hipertensão 3 >140/<90 Hipertensão sistólica isolada
05	Colesterol	Indica o nível de colesterol no sangue do paciente na hora da consulta.	> 45 Bom (HDL) > 110 Ruim (LDL)
06	Diabetes	Indica a presença ou não da condição que resulta em alta do nível de açúcar no sangue do paciente na hora da consulta.	0 - Falso 1 - Verdadeiro
07	Electro Cardio Repouso	Indica a atividade elétrica do coração quando o paciente está em repouso.	0 - Normal 1 - Não normal 2 - Alta
08	Frequência Cardíaca Máximo	Indica o número de batimentos cardíacos por unidade de tempo.	100 - 170 - Normal ≥ 170 - Alta
09	Angina Induzida	Indica a presença ou não de dor no peito do paciente na hora da consulta.	0 - Negativo 1 - Positivo
10	Depressão induzida	Indica a presença de depressão induzida por medicamentos do paciente	0 - Negativo > 0 - Positivo
11	Exercício De Pico	Indica se o paciente faz exercícios com alta intensidade	0 - Ascendente 1 - Plana 2 - Descendente
12	Estado de Vasos	Indica se as veias estão no estado boas ou não	0 = Bom 1 = Ruim
13	Talassemia	Indica a presença ou o tipo de talassemia no sangue.	1 - Normal 2 - Alfa 3 - Beta
14	Estado de hipertensão arterial	Indica se o paciente apresenta hipertensão arterial ou não	0 - Negativo 1 - Positivo

Fonte: Cleveland Heart Disease (2020)

### 3.4 MODELOS DE CLASSIFICAÇÃO

Uma vez que o conjunto de dados foi tratado, definiu-se os classificadores, a fim de prever com precisão as classes dos novos exemplos. O critério para selecionar os algoritmos deu-se pela ampla utilização em projetos de pesquisa e alta taxa de acerto. Primeiramente, usou-se a técnica de classificação Rede Bayesiana. Logo após o *LogitBoost*, o *Multi-layer Perceptron* e, por fim, o classificador J48.

É importante mencionar que, como todos os métodos são probabilísticos, a literatura considera o número 30 como o valor médio para definir o “*Random seed for XVal*”, gerador de números aleatórios de divisão de dados, isto é, o modelo é executado 30 vezes alterando essa semente e, com isso, alterando o desempenho do algoritmo. No final, calcula-se a média dos resultados de cada método.

A *seed*, chamada de semente aleatória ou vetor, tem como objetivo iniciar o algoritmo com um número pseudoaleatório na hora da execução. É necessário que o modelo tenha um número da semente definida antes da sua aplicação, caso contrário, ele gera uma sequência de números independentes um dos outros que pode resultar com diferentes respostas na saída (PARK, 1988).

Também foi usada a abordagem de validação cruzada (em inglês, *cross validation* - cv). Isto significa que, para cada *seed*, o método executa utilizando diferentes porções de base (9/10 e 1/10) onde cada porção passará no processo interno.

Em seguida, a fase do treinamento do modelo e validação são aplicadas para todos dados disponíveis na base, fazendo o treinamento repetidamente “k” vezes, isto é, o modelo selecionado será executado dez vezes e 90% dos dados oferecidos são utilizados como grupo de treinamento e os outros 10% restantes como grupo de teste que é a segunda porção reservada para avaliar a melhor solução.

## 4 RESULTADOS E DISCUSSÃO

Neste capítulo são apresentados e discutidos os resultados dos experimentos realizados neste trabalho.

### 4.1 EXPERIMENTO 1

Para realização do primeiro experimento, utilizou-se o algoritmo Rede Bayesiana. Observou-se que, considerando a configuração padrão dos parâmetros, o algoritmo obteve as melhores taxas de acerto. Como resultado deste experimento tem-se 82,47% de acurácia, obtidos da média das 30 execuções realizadas.

Quanto à sensibilidade e a especificidade, duas variáveis importantes para mensurar a precisão do diagnóstico, o modelo Rede Bayesiana obteve 83,68% de sensibilidade, isto é, o algoritmo classificou de uma forma correta os pacientes hipertensos, e a especificidade de 82,69%, que representa os pacientes da classe negativo classificados de forma correta.

Em seguida, alterou-se apenas a configuração *useKernelEstimator* para *true* de modo a observar o comportamento do algoritmo e buscar melhor acurácia. Optou-se pela alteração deste parâmetro para *true* pois o mesmo é indicado para atributos numéricos, o que é o caso da base de dados utilizado. Observou-se o seguinte resultado: 82,50% de acurácia, isto é, um aumento de 0,03% com relação ao primeiro teste realizado. Obteve-se também 83,64% de sensibilidade e 82,62% de especificidade. A Tabela 4 apresenta os resultados do experimento.

**Tabela 4 - Acurácia do algoritmo Rede Bayesiana Experimento 1**

Modelo	Métricas		
	Sensibilidade (%)	Especificidade (%)	Acurácia (%)
Rede Bayesiana	83,68	82,69	82,47
	83,64	82,62	82,50

Fonte: O Autor (2020)

Os resultados apresentados na tabela 4 mostram o desempenho do modelo durante o experimento, percebe-se que no segundo teste houve aumento na taxa de acurácia que é considerado no contexto da pesquisa e uma pequena diminuição nas taxas da sensibilidade e especificidade.



## 4.2 EXPERIMENTO 2

Para realização do segundo experimento, usou-se o modelo LogitBoost (LB) com as configurações padrão do Weka, os resultados obtidos foram: 81,23% na média, das instâncias classificadas corretamente e 18,77% das instâncias classificadas incorretamente. Quanto à sensibilidade que é a proporção dos pacientes doentes classificados como positivos obteve-se 80,27% e a taxa da especificidade, pacientes sem hipertensão classificados de maneira correta, 81,32%.

Na busca por melhores resultados alterou-se o parâmetro *useResampling* para *True*. Com esta alteração, foi obtido um resultado 81,32%, um aumento de 0,10% da acurácia. Com essa configuração, o modelo LogitBoost alcançou 80,56% de sensibilidade e 81,59% de especificidade.

Quando o atributo *useResampling* é configurado para *True*, ele faz com que o algoritmo use a reamostragem que é a ponderação na hora do treinamento. Quando o atributo é *falso*, usa-se a ponderação a cada iteração do processo de treinamento e um peso é atribuído a cada amostra no conjunto de treinamento.

Esse reajuste na configuração do modelo resulta em um desempenho melhor durante como é mostrado na Tabela 5.

**Tabela 5 - Acurácia do algoritmo LogitBoost Experimento 2**

Modelo	Métricas		
	Sensibilidade (%)	Especificidade (%)	Acurácia (%)
LogitBoots	80,27	81,32	81,23
	80,56	81,59	81,32

Fonte: O Autor (2020)

## 4.3 EXPERIMENTO 3

Para realização do terceiro experimento, usou-se o modelo *Multi-layer Perceptron* (MLP) com a configuração padrão de Weka. O modelo alcançou uma taxa de 82% das instâncias classificadas corretamente e 18% das instâncias classificadas incorretamente. E a taxa média da sensibilidade foi de 81,7% e 80,8% de especificidade.

Seguidamente, alterou-se as configurações do modelo com objetivo de melhorar a eficácia do modelo, os parâmetros alterados foram *decay*, para o valor

*True*, esse atributo fará com que a taxa de treinamento diminua. Isso dividirá a taxa de aprendizado inicial pelo número da época “o círculo de cada treinamento” para determinar qual deve ser a taxa de aprendizado atual, isso pode ajudar a impedir que a rede se desvie da saída de destino, além de melhorar o desempenho geral; o *hiddenlayers* (quantidade de camadas de neurônios ocultas) foi alterado para 2, o valor padrão é “a” (número 0); e o *seed*, que é a semente usada para inicializar o gerador de números aleatórios, são usados para definir os pesos iniciais das conexões entre os nós e também para misturar dados de treinamento foi trocado para 1.

Com essa configuração, os resultados foram: 81,95% de acurácia, teve uma queda de 0,05% comparado com o teste anterior, um valor estatisticamente considerado no desempenho do modelo. 83,06% de sensibilidade e 81,10% de especificidade.

A Tabela 6 apresenta as medidas de desempenho do algoritmo obtidas neste experimento.

**Tabela 6 - Acurácia do algoritmo *Multi-layer Perceptron* Experimento 3**

Modelo	Métricas		
	Sensibilidade (%)	Especificidade (%)	Acurácia (%)
Multi-layer Perceptron	81,70	80,80	82,00
	83,06	81,10	81,95

Fonte: O Autor (2020)

É levado em consideração o primeiro resultado do experimento com a taxa de 82% de acurácia, apesar de que as duas taxas sensibilidade e especificidade aumentaram após os ajustes feitos na configuração do modelo.

#### 4.4 EXPERIMENTO 4

Para realização do quarto experimento, utilizou-se o algoritmo J48 (C4.5), baseado em árvore de decisão, com as configurações padrão do Weka. O resultado alcançado é de 77% instâncias classificadas corretamente e 23% instâncias classificadas incorretamente a taxa da Sensibilidade 76,85% e a taxa da Especificidade 82,7%.

Em seguida, decidiu-se alterar o atributo *reducedErrorPrunin* na configuração do modelo com intuito de atingir a melhor performance do modelo. Esse atributo foi

trocado para *True*, isto é, ele visa reduzir os erros que a árvore pode gerar durante o treinamento por meio da “poda” da árvore.

O resultado obtido neste experimento é de 77,18% das instâncias classificadas corretamente e 23,00% das instâncias classificadas incorretamente. A média da taxa da sensibilidade é de 75,88% e a especificidade foi de 78,21%, duas variáveis importantes para medir a precisão do diagnóstico.

A Tabela 7 apresenta detalhes da acurácia obtida neste experimento. No segundo experimento o modelo teve a sua taxa média de 77,18% com o aumento de 0,18% comparado com o resultado anterior antes do ajuste.

**Tabela 7 - Acurácia do algoritmo J48 Experimento 4**

Modelo	Métricas		
	Sensibilidade (%)	Especificidade (%)	Acurácia (%)
J48	76,85	82,70	77,00
	75,88	78,21	77,18

Fonte: O Autor (2020)

#### 4.5 DISCUSSÃO

A Tabela 8 apresenta o resumo dos resultados da classificação. A principal métrica utilizada para otimização de cada modelo durante o período de aprendizado foi *acurácia*, sem esquecer da *sensibilidade* e *especificidade*. Finalmente, a avaliação de desempenho dos modelos selecionados demonstrou desempenho satisfatório, com *Acurácia* superior a 75% para todos os modelos. É possível observar que as acurácias geradas dos classificadores na maior parte dos experimentos encontram-se na faixa de 77 a 83%, e os melhores resultados foram encontrados para os modelos Rede Bayesiana (82,5% de acerto) e *Multi-layer Perceptron* (82,00% de acerto) com pouca diferença os dois modelos se destacaram.

Dentre os modelos aplicados, a Rede Bayesiana destacou-se com o acerto de mais de 82,5%. Esse modelo é de raciocínio probabilístico, isto é, trabalha bem em ambientes que existem informações incompletas ou informações aproximadas e com relação ao contexto dessa pesquisa a teoria da probabilidade com enfoque Bayesiano teve o seu desempenho melhor que outros modelos, pois ela considerou a probabilidade como o grau de certeza da eventualidade de um paciente ser hipertenso ou não.

Também, deve-se levar em consideração outra análise que é em relação às taxas de sensibilidade e especificidade. Um modelo que apresenta uma taxa de sensibilidade que se aproxima de 100%, se caracteriza como sendo um bom modelo para classificar os casos positivos corretamente (como hipertenso). A Rede Bayesiana destacou-se com a taxa de sensibilidade comparado aos demais modelos utilizados, cerca de 83,69%. Assim como a acurácia, a taxa de especificidade obtida (82,70%) é considerada relevante. Quanto menor a taxa de especificidade maior é a taxa de falso positivo, ou seja, o modelo pode classificar um paciente como hipertenso incorretamente, causando desconforto para o paciente. Por fim, observa-se que o J48 obteve a menor taxa de sensibilidade, isto significa que o modelo tem dificuldade em classificar os casos positivos corretamente.

**Tabela 8 - Resultados de classificação final dos quatros modelos**

Modelos	Sensibilidade (%)	Especificidade (%)	Acurácia (%)
Rede Bayesiana	83,64	82,62	<b>82,50</b>
LogitBoost	80,56	81,59	81,32
Multi-layes Perceptron	81,70	80,80	<b>82,00</b>
J48	75,88	78,21	77,18

Fonte: O Autor (2020)

A Tabela 9 apresenta a comparação dos resultados obtidos neste trabalho e os trabalhos relacionados. Observa-se que o trabalho de Shubhankar (2019) citado na seção 2.1.2, utilizou a mesma base de dados *Cleveland Heart Disease*, porém utilizou-se *Python* como linguagem de programação. Ele obteve com o modelo *Logistic Regression* a taxa de acerto de 80.32%. Já Oliveira et al. (2017) obtiveram os melhores resultados com os modelos Redes Neurais Artificiais e Regressão Logística que atingiram 75.24% e 74.98% de acerto, respectivamente.

Observa-se que o modelo Rede Bayesiana atingiu 82,47% de acerto, pouco acima da acurácia obtida por Shubhankar (2019). Pode-se supor que a diferença na configuração dos algoritmos, as ferramentas utilizadas para classificação, o pré-processamento dos dados e a seleção dos atributos tenham influenciado nos diferentes resultados obtidos.

**Tabela 9 - Comparação dos resultados dos trabalhos relacionados**

Trabalho	Modelos	Base de dados	Atributos	Instâncias	Acurácia
Shubhankar (2019)	Logistic Regression	Cleveland Heart Disease	14	303	80,00%
Oliveira et al., (2017)	Redes Neurais Artificiais	Elsa (Brasil)	24	403	75,00%
O Autor (2020)	Rede Bayesiana	Cleveland Heart Disease	14	303	82,50%

Fonte: O autor (2020)

#### 4.6 COMPARAÇÃO DOS MODELOS

Nesta seção são apresentados os resultados dos testes realizados para a comparação dos modelos de aprendizagem de máquina usados neste trabalho.

Aplicou-se o teste estatístico Friedman para comparar os modelos usados na pesquisa, **Teste Friedman**, teste não paramétrico realizado no IBM SPSS, com objetivo de encontrar a diferença entre modelos nos experimentos realizados. O Quadro 2 mostra cada modelo com o valor médio calculado durante a execução.

Quadro 2 - Classificação dos modelos

	Postos de média
Rede Bayesiana	1,48
LogistBoot	2,26
MLP	1,95
J48	3,95

Fonte: O Autor (2020)

De acordo com os resultados obtidos na Quadro 1 de classificação acima, no qual, apresenta a média padrão dos modelos, percebe-se que a Rede Bayesiana se difere dos modelos MLP, LogitBoost e J48, alcançando a média de 1.48 que é menor do valor convencional 1.64, e é considerado como o melhor modelo a ser aplicado nesse contexto de diagnóstico de hipertensão arterial.

Durante os experimentos, cada modelo foi aplicado 30 vezes de acordo com o número de *seed* definido antes da execução, o teste de Friedman está sendo aplicado para ver o efeito de cada *seed* no desempenho dos modelos. Na Tabela 10 apresenta

a estatística descritiva dos modelos, a mediana, pois o Friedman não se aplica no conjunto de dados onde a média não é uma boa representação, o valor erro desvio etc.

**Tabela 10 - Estatísticas descritivas**

	N	Média	Erro Desvio	Mínimo	Máximo	Percentis		
						25º	50º(mediana)	75º
Rede Bayesiana	30	1,45	0,562	1	3	1,00	1,00	2,00
LogitBoost	30	2,58	0,720	1	4	2,00	3,00	3,00
MLP	30	1,93	0,785	1	3	1,00	2,00	3,00
J48	30	3,39	0,365	2	4	4,00	4,00	4,00

Fonte: O Autor (2020)

Como a estatística se preocupa com a variabilidade dos dados amostrais, a Tabela 11 apresenta diferentes valores das amostras analisadas.

**Tabela 11 - Amostragens**

Amostra1	Amostra 2	Estatística de Teste	Std. Erro	Erro Estatística de Teste	Sig.	Sig. Aj.
Rede Bayesiana	MLP	-0,467	0,333	-1,400	0,162	0,969
Rede Bayesiana	LogitBoost	-1,133	0,333	-3,400	0,001	<b>0,004</b>
Rede Bayesiana	J48	-2,467	0,333	-7,400	0,000	<b>0,000</b>
MLP	LogitBoost	0,667	0,333	2,000	0,046	0,273
MLP	J48	-2,000	0,333	-6,000	0,000	<b>0,000</b>
LogitBoost	J48	-1,333	0,333	-4,000	0,000	<b>0,000</b>

Fonte: O Autor (2020)

## 5 CONSIDERAÇÕES FINAIS E TRABALHOS FUTUROS

O presente trabalho teve como objetivo principal aplicação e comparação dos modelos de aprendizagem de máquina no diagnóstico de hipertensão arterial, com a finalidade de prever a ocorrência da doença, para isso, utilizou-se dados disponibilizados no repositório *Cleveland Clinic*.

Tem-se observado que a aplicação de modelos de aprendizado de máquina no diagnóstico de hipertensão arterial pode produzir resultados mais precisos que uma avaliação feita por profissionais da área de saúde, que são suscetíveis a erros devido a diversos fatores. Com isso, a utilização desses modelos no diagnóstico de hipertensão é necessária a fim de facilitar a tomada de decisão para evitar riscos causados pela doença e até mesmo a perda da vida.

Considerando os resultados obtidos nos experimentos realizados neste trabalho, de modo geral, mostrou-se que todos os modelos produziram resultados considerados satisfatórios com acerto acima de 75%, demonstrando assim a viabilidade de detectar pacientes com hipertensão através de dados clínicos.

Com base no presente trabalho, indica-se algumas recomendações para a continuação da pesquisa nesta área. Alguns pontos merecem aprofundamento em pesquisas ou trabalhos futuros. Destes, os principais são:

- a) Utilização de outros modelos não aplicados neste trabalho, como por exemplo, Clusterização;
- b) A utilização de outras bases de dados como, por exemplo, do estado da Paraíba;
- c) Criação de uma ferramenta de auxílio ao diagnóstico de fácil utilização por parte dos profissionais de saúde, os quais podem fazer uso da ferramenta no momento da consulta, aplicando os dados coletados dos pacientes, apoiando o diagnóstico e diminuindo a incerteza no diagnóstico.

## REFERÊNCIAS BIBLIOGRÁFICAS

APPOLINÁRIO, F. **Dicionário de metodologia científica**: um guia para a produção de conhecimento científico. São Paulo: Atlas, 2004.

AUSTIN, **Inteligência artificial**: Machine Learning. 6. ed. São Paulo: [s.n.], 2017. p.119-121.

BRINK, H; RICHARDS, J W. **Real-World Machine Learning**. Shelter Island. NY: Manning, 2013. Disponível em: [http://www.manning.com/brink/RWML\\_MEAP\\_CH01.pdf](http://www.manning.com/brink/RWML_MEAP_CH01.pdf). Acesso em: 14 nov. 2019.

CERA, M.C. **Uso de Redes Neurais para o Reconhecimento de Padrões**. Disponível em: <http://www.inf.ufrgs.br/procpar/disc/cmp135/trabs/mccera/t1/padrões.pdf>. Acesso em: novembro, 2019.

FACELI; KATTI LORENA; ANA CAROLINA; CARVALHO. **Inteligência Artificial**: Uma Abordagem de Aprendizagem de Máquina. Rio de Janeiro: LTC, 2011.

FRIEDMAN, J. H.; HASTIE, T.; TIBSHIRANI, R. **Additive Logistic Regression**: A statistical view of boosting. *The Annals of Statistics*, v. 28, p. 337 à 407, 2000.

FURNKRANZ, J.; GAMBERGER, D.; LAVRAC, N. **Foundations of Rule Learning**. [S.L.]: Springer-Verlag Berlin, 2012.

HALL, M.; FRANK, E.; HOLMES, G.; PFAHRINGER, B.; REUTEMANN, P.; WITTEN, I. **The Weka data mining software**: an update. *ACM SIGKDD Explorations Newsletter*, ACM, v. 11, n. 1, p. 10–18, 2009.

HAYKIN, S. **Redes Neurais**: Princípios e Práticas. BOOKMAN, São Paulo, 2ª ed. 2001.

HOSSIN, M; SULAIMAN M. N. A. **Review on evaluation metrics for data classification evaluations**. *Em International Journal of Data Mining & Knowledge Management Process*, v. 5, n. 2, p. 1, 2015.

INSTITUTO DE CARDIOLOGIA DO LAGO, **Mapa de Pressão Arterial**: Para que serve? Disponível em: <https://telemedicinamorsch.com.br/blog/o-que-e-mapa-de-pressao-arterial>. 24 fevereiro, 2020.

INSTITUTO DE CARDIOLOGIA DE LAGO, **MAPA**: Monitorização Ambulatorial da Pressão Arterial. Disponível em: <https://cardiolago.com.br/2019/09/06/mapa-significa-monitorizacao-ambulatorial-da-presso-arterial-feita-geralmente-a-cada-20/>. Acesso em: fevereiro, 2020.

NOGUEIRA PC & CALIRI MHL. **Fatores de Risco para Úlcera de Pressão em pacientes com trauma medular**. Análise da literatura científica nacional. 6º Simpósio de Iniciação Científica da USP, Ribeirão Preto, novembro de 1998.



NUNES, S; ANDRADE, R; HERNANE, S. **Redes Neurais Artificiais**. 2ed. Para engenharia e ciências aplicadas. Universidade de São Paulo. 2016.

OLIVEIRA; RODRIGUES, A. **Comparação de algoritmos de aprendizagem de máquina para construir um modelo preditivo para detecção de diabetes não diagnosticada**: Elsa-Brasil: Estudo de acurácia. São Paulo Medical Journal, v. 135, n. 3, p. 234-246, jun 2017.

PARK, S; MILLER, KW. **Random number generators**: good ones are hard to find, communications of the ACM, v.31 n.10, p.1192-1201, Oct.1988.

ROSS QUINLAN, C4.5: **Programs for Machine Learning**. Morgan Kaufmann Publishers, San Mateo, CA. 1993.

RUSSEL; NORVIG; PETER. **Artificial Intelligence**. 2. Ed. Rio de Janeiro: Campus, 2004.

SHUBHANKAR, R. **Heart Disease Prediction**. Disponível em: <https://towardsdatascience.com/heart-disease-prediction-73468d630cfc>. Acesso em: setembro, 2019.

SOCIEDADE BRASILEIRA DE CARDIOLOGIA. **VII Diretrizes de Monitorização Ambulatorial da Pressão Arterial (MAPA) e Monitorização Residencial de Pressão Arterial (MRPA)**. Arquivos Brasileiros de Cardiologia, São Paulo, v.107, p. 1-32, 3 setembros, 2019.