Comparação de Algoritmos de Aprendizado de Máquina na Classificação de Neoplasias Mamárias

Tulio Araujo Santos De Oliveira

Universidade Federal de Uberlândia
Faculdade de Computação
Bacharelado em Sistemas de Informação

Monte Carmelo - MG
2021

Tulio Araujo Santos De Oliveira

Comparação de Algoritmos de Aprendizado de Máquina na Classificação de Neoplasias Mamárias

Trabalho de Conclusão de Curso apresentado à Faculdade de Computação da Universidade Federal de Uberlândia, Minas Gerais, como requisito exigido parcial à obtenção do grau de Bacharel em Sistemas de Informação.

Área de concentração: Ciência da Computação

Orientadora: Dra. Fernanda Maria da Cunha Santos

Monte Carmelo - MG
2021

Este trabalho é dedicado à minha mãe pois é graças ao seu esforço que hoje posso concluir o meu curso.

Agradecimentos

Agradeço primeiramente a Deus e a minha família por sempre terem me apoiado nessa jornada que foi se graduar na UFU. Agradeço a todos os meus amigos que fizeram parte desta jornada. Agradeço minha professora orientadora Fernanda, por ter aceitado acompanhar-me neste trabalho, seu empenho e paciência em me ajudar e me proporcionar tanto aprendizado e crescimento. Meu muito obrigado a todos!

Resumo

O câncer de mama no Brasil e no mundo é um problema relevante de saúde pública, o que justifica as intensas pesquisas e esforços para aumentar a expectativa de vida dos pacientes. O fator crucial para uma maior expectativa de vida dos pacientes é a detecção precoce da doença. Assim, a criação de procedimentos computacionais aliados às metodologias tradicionais da Medicina podem contribuir significativamente para o diagnóstico de células cancerígenas. Então, este trabalho propõe um classificador constituído pela Análise de Componentes Principais e por técnicas de aprendizado de máquina com a finalidade de predizer as neoplasias mamárias. Os métodos de aprendizado de máquina escolhidos foram Redes Neurais Artificiais, *Support Vector Machine*, Naive Bayes e Árvore de Decisão. A base de dados utilizada foi *Wisconsin Diagnostic Breast Cancer*, e os resultados apresentaram um desempenho satisfatório para os classificadores constituídos pelas Redes Neurais Artificiais e pelo *Support Vector Machine*, apresentando acima de 90% de acertamentos para as métricas de avaliação acurácia, precisão, revocação, medida-F1 e área ROC.

Palavras-chave: Aprendizado de Máquinas, Classificação de padrões, Câncer de mama, *Support Vector Machine*, Naive Bayes.

Lista de ilustrações

Lista de siglas

AM Aprendizado de máquina

CAD Diagnóstico Assistido por Computador

MLP Multi Layer Perceptron

PCA Principal Component Analysis

RNAs Redes Neurais Artificiais

ROC Receiver Operating Characteristic

SVM Support Vector Machine

Sumário

Capítulo 1

Introdução

1.1 Motivação

O câncer de mama é hoje um relevante problema de saúde pública. É considerada a neoplasia maligna mais incidente em mulheres na maior parte do mundo (INCA, 2019). De acordo com as últimas estatísticas mundiais do Globocan 2018 (BRAY et al., 2018), foram estimados 2,1 milhões de casos novos de câncer e 627 mil óbitos pela doença. No Brasil, as estimativas de incidência de câncer de mama para o ano de 2019 são de 59.700 casos novos, o que representa 29,5% dos cânceres em mulheres, atrás apenas do câncer de pele não melanoma. Em 2016, ocorreram 16.069 mortes de mulheres por câncer de mama no país (INCA, 2019).

A forma mais eficaz na cura desta doença é a detecção precoce. A mamografia é uma das melhores técnicas para o rastreamento do câncer de mama disponível atualmente, capaz de registrar imagens da mama com a finalidade de diagnosticar a presença ou ausência de estruturas que possam indicar a doença. Com esse tipo de exame pode-se detectar o tumor antes que ele se torne palpável. O exame de mamografia é o diagnóstico definido por um radiologista para diagnosticar o câncer de mama nas predições primárias da doença, requer experiência e muita atenção, o que implica em uma grande limitação.

Uma ferramenta alternativa que auxilia os radiologistas na leitura de imagens e diagnóstico precoce do câncer de mama é o Diagnóstico Assistido por Computador (CAD). Os sistemas CAD não substituem o especialista, pelo contrário, são úteis para diminuir a incerteza no diagnóstico. Os sistemas CAD podem dispor de técnicas para melhorar a qualidade da imagem e, consequentemente, a visualização e localização de lesões suspeitas, extrair características a partir de imagens e classificar os achados de acordo com sua probabilidade de malignidade (SOUTO, 2013).

Pesquisadores da saúde pública e da comunidade científica tecnológica, estão fazendo uso de técnicas de Aprendizado de máquina (AM) agregados aos sistemas computacionais de diagnósticos, com o intuito de melhorar as taxas de detecção precoce do câncer de mama. Contudo, existem tradicionais sistemas computacionais de diagnóstico em

Medicina que não processam grandes quantidades de dados, por isso é importante ter ferramentas computacionais atuando como sistemas de apoio à decisão no processamento de registros dos pacientes (FIDA et al., 2011). Várias técnicas de AM foram propostas para projetar sistemas de classificação precisos para vários problemas médicos (Yungang Zhang et al., 2012).

A escolha da técnica de classificação mais precisa entre vários classificadores é um dos problemas na seleção do modelo. Detectar o melhor nível de complexidade ou selecionar o classificador mais preciso é o objetivo das abordagens de seleção de modelo. Além disso, os melhores classificadores são escolhidos com base no desempenho, custos de software e preferências do usuário. Contudo, há um problema em obter resultados mais precisos em uma quantidade limitada de dados, o que pode ser melhorado usando o pré-processamento. Técnicas de pré-processamento são comumente aplicadas no conjunto de dados antes da etapa de classificação.

1.2 Objetivos e Desafios da Pesquisa

1.2.1 Objetivo Geral

O objetivo deste trabalho é criar um classificador de padrões, constituído por algoritmos de Aprendizado de Máquina e pela Análise Componentes Principais (PCA), com o propósito de analisar e avaliar a base de dados Wisconsin e, por fim, classificar as neoplasias mamárias. Devido à diferentes algoritmos presentes em AM, optou-se por testar quatro possíveis e, posteriormente, aferir qual possui melhor desempenho sob os dados em estudo. A base de dados contém atributos quantitativos extraídos das mamografias, além de parâmetros estatísticos calculados a partir dos próprios atributos, visando a classificação da citologia entre maligno ou benigno.

1.2.2 Objetivos Específicos

- ❑ Estudo da técnica de pré-processamento Principal Component Analysis (PCA) que auxilia o agrupamentos dos atributos para obtenção do vetor de características;
- ❑ Uso de bibliotecas para implementação de algoritmos de Aprendizado de Máquina para a classificação de células mamárias entre malignas ou benignas utilizando a linguagem Python. Dentre as técnicas de AM destacam-se as Redes Neurais Artificiais (RNAs), Árvore de Decisão, Naive Bayes, Support Vector Machine (SVM);
- ❑ Análise e avaliação das características e do desempenho dos classificadores de Aprendizado de Máquinas implementados neste estudo.

1.3 Contribuições

Definição de um classificador de padrões, com a função de auxiliar exames mamográficos, implementando as técnicas de AM: RNAs usando a arquitetura Multi Layer Perceptron (MLP), o algoritmo Naive Bayes, Árvore de Decisão e o SVM. Ademais, fazer uma análise e comparação dos resultados obtidos com outros que já foram descritos em trabalhos referenciados, pois existem várias pesquisas sobre câncer de mama usando o conjunto de dados de câncer de mama de Wisconsin.

1.4 Organização da Monografia

Este trabalho encontra-se organizado em 4 capítulos, a saber: o Capítulo 2 apresenta um embasamento teórico da técnica de PCA, dos algoritmos de AM, além de descrever os principais trabalhos que auxiliaram a fundamentação deste estudo. O Capítulo 3 destaca os algoritmos de AM implementados em Python, a descrição das informações contidas na base de dados e a explicação de todos os resultados alcançados. No Capítulo 4, as conclusões, críticas e sugestões de trabalhos futuros são apresentados.

Capítulo 2

Fundamentação Teórica

Neste capítulo são apresentados os principais conceitos relacionados com o tema deste trabalho. São eles: redes neurais MLP, algoritmo Naive Bayes, Support Vector Machine, Árvore de Decisão, PCA e métodos de avaliação do desempenho dos algoritmos. Ademais, será descrito os artigos que contribuíram como referencial teórico para a evolução deste estudo.

2.1 Redes Neurais Artificiais

Redes neurais artificiais são modelos computacionais definidos no âmbito de AM e que possuem sua inspiração nos neurônios do cérebro humano e em suas conexões, as quais possibilitam a ligação dos neurônios e o fluxo do conhecimento. Os neurônios artificiais recebem os sinais de entrada, representado por vetores contendo dados de entrada, e realizam uma função matemática simples, gerando um sinal de saída, que é representado por um ou mais valores. O resultado final consiste em um modelo capaz de moldar e aprender a executar determinadas tarefas específicas sobre o ajuste correto dos parâmetros, representado por vetores ou matrizes de pesos sinápticos. (HAYKIN, 2007).

As RNAs são aplicadas em diversas áreas de atuação com as seguintes funções (SILVA; SPATTI; FLAUZINO, 2010):

- ❏ Aproximador universal de funções;
- ❏ Controle de processo;
- ❏ Reconhecimento e classificação de padrões;
- ❏ Agrupamento de dados (clusterização);
- ❏ Sistemas de previsão;
- ❏ Otimização de sistemas;

- ❏ Memórias associativas.

A arquitetura de uma rede neural artificial define a forma como os seus diversos neurônios estão arranjados. A rede Perceptron é uma forma mais simples de configuração de uma rede neural artificial, idealizada por Rosenblatt (1958), objetivando-se identificar padrões geométricos (SILVA; SPATTI; FLAUZINO, 2010).

O perceptron consiste de dois tipos de nodos: nodos de entrada que são usados para representar os atributos de entrada e um nodo de saída, que é usado para representar a saída do modelo. Os nodos de uma arquitetura de rede neural são conhecidos comumente como neurônios ou unidades. Em um perceptron, cada nodo de entrada é conectado através de uma ligação ponderada com o nodo de saída. A ligação ponderada é usada para emular a força da conexão sináptica entre neurônios, da mesma forma que em sistemas neurais biológicos, treinar um modelo perceptron requer que se adaptem os pesos das ligações até que se apropriem aos relacionamentos entrada-saída dos dados correspondentes (TAN; STEINBACH; KUMAR, 2009).

A Figura 1 apresenta um neurônio de uma rede Perceptron.

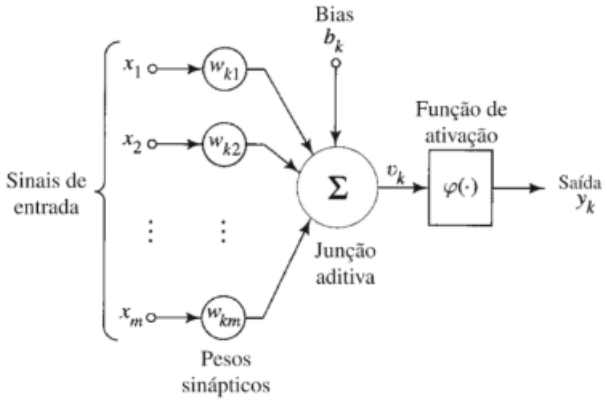

Figura 1 – Modelo não linear de um neurônio. Extraído de (HAYKIN, 2007)

Podemos identificar os três elementos básicos do neurônio:

- ❏ As sinapses, cada uma caracterizada por um peso ou uma força.
- ❏ O somador, que soma os sinais de entrada ponderados pelas sinapses do neurônio (combinador linear).
- ❏ A função de ativação (ou função restritiva) que delimita o intervalo de saída do neurônio.

2.1.1 Redes Perceptron Multicamadas

As redes Perceptron Multicamadas (MLP) possuem uma estrutura mais complexa do que a do modelo perceptron. A rede pode conter diversas camadas intermediárias

entre suas camadas de entrada e de saída. Tais camadas intermediárias são chamadas de camadas ocultas e os nodos internos nestas camadas são chamados nodos ocultos. A estrutura resultante é conhecida como rede neural multicamadas (Veja a Figura 2). Esse tipo de arquitetura é denominado redes *feedforward* (alimentação para frente) de camadas múltiplas, cujos os nodos de uma camada estão conectados apenas aos nodos da próxima camada.

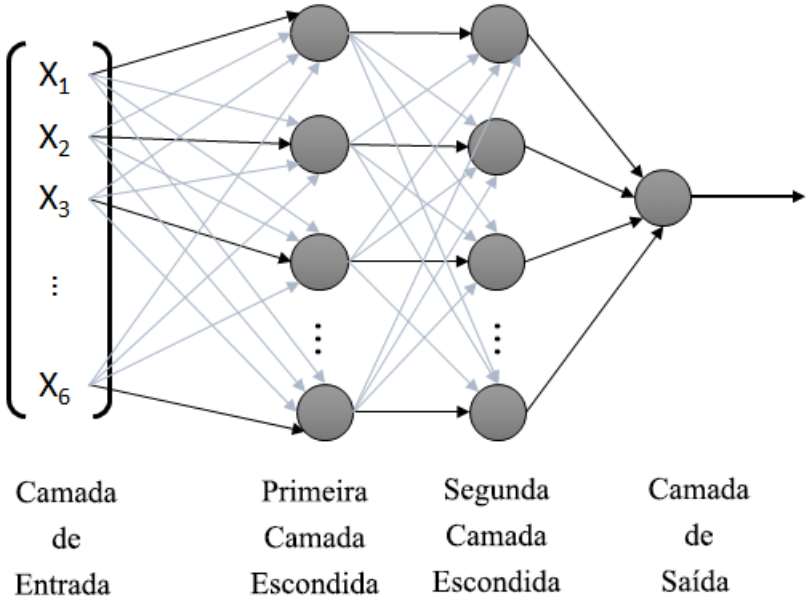

Figura 2 – Exemplo da arquitetura de uma rede neural artificial multicamadas com alimentação para frente (VIDAL et al., 2015).

Um dos destaques em RNAs está na capacidade de aprender a partir de um conjunto de amostras que exprimem o comportamento do sistema. Ou seja, a rede neural irá aprender o relacionamento entre um conjunto de dados de entrada e suas respectivas saídas e, em seguida, será capaz de generalizar soluções. A fase de aprendizado é o que se chama de processo de treinamento de uma rede neural. Esse processo define algoritmos de aprendizagem que utilizam de funções matemáticas, que são denominadas funções de ativação, para balancear os pesos sinápticos e limiares dos neurônios.

A rede pode usar diferentes tipos de funções de ativação, como a função sinal, tangentes hiperbólicas, sigmoides (Logísticas) e lineares, conforme mostrado na Figura 3. Estas funções de ativação permitem que os nodos ocultos e de saída produzam valores de saída que sejam não lineares nos seus parâmetros de entrada.

O ajuste dos pesos e do limiar de cada um dos neurônios da rede MLP são efetuados utilizando o processo de treinamento supervisionado, isto é, para cada amostra dos dados de entrada obtém a respectiva saída (resposta) desejada. O algoritmo de aprendizado aplicado no decorrer do processo de treinamento é denominado *backpropagation* ou algoritmo de retropropagação do erro (SILVA; SPATTI; FLAUZINO, 2010).

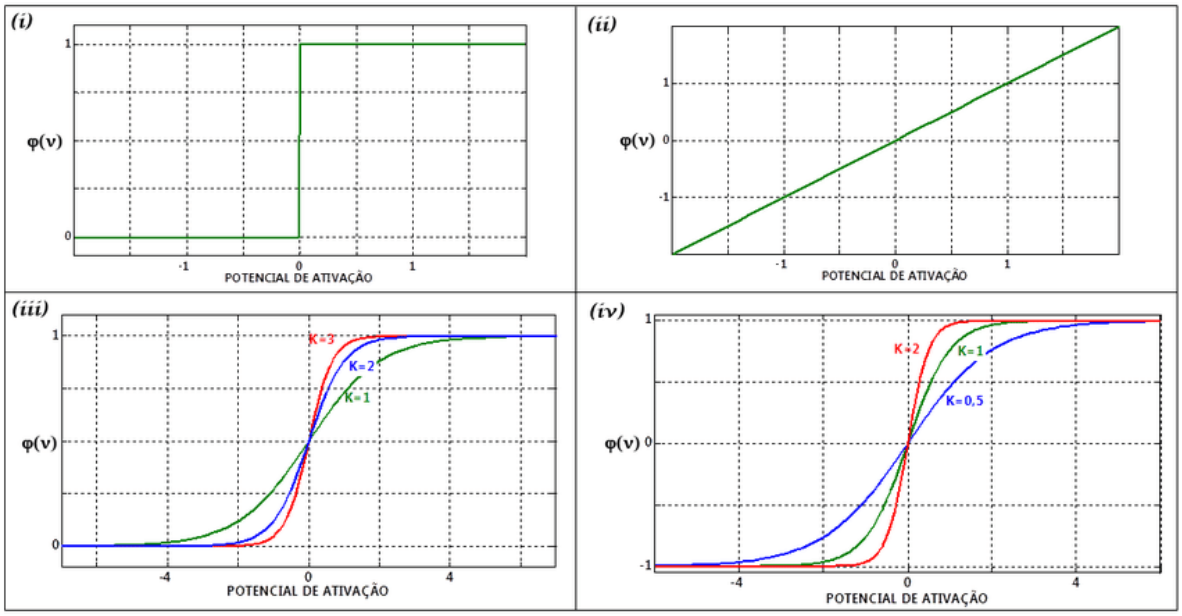

Figura 3 – Tipos de funções de ativação em redes neurais artificiais (OLIVEIRA, 2013).

2.2 Naive Bayes

O algoritmo Naive Bayes é um classificador estatístico baseado no Teorema de Bayes, que prediz a probabilidade de um determinado dado pertencer a uma classe em particular. As probabilidades são estimadas de acordo com a frequência de cada valor para os registros de treino. Assim, dada uma nova instância, o classificador faz a estimativa de probabilidade de o registro feito pertencer a uma nova classe específica, considerando que os atributos são condicionalmente independentes (BERTON, 2011).

O teorema de Bayes é definido do seguinte modo: seja $C = \{c1, c2, ..., c_l\}$ o conjunto de classes dos dados e x uma instância de classe desconhecida. Considerando que x pertença a uma das classes do conjunto C, deseja-se determinar $P(c_i|x), 1 \leq i \leq l$, ou seja, a probabilidade da classe c_i dada a instância x. O cálculo da probabilidade a posteriori da classe c_i condicionada a x, $P(c_i|x)$, é dado pela regra de Bayes:

$$P(c_i|x) = \frac{P(x|c_i)P(c_i)}{P(x)} \tag{1}$$

onde $P(c_i)$ é a probabilidade *a priori* da classe c_i , $P(x)$ é a probabilidade *a priori* de x e $P(x|c_i)$ é a probabilidade *a posteriori* de x condicionada a classe c_i. As probabilidades $P(c_i)$, $P(x)$ e $P(x|c_i)$ são estimadas a partir das instâncias de treinamento.

O classificador Bayesiano parte do ponto onde eventos possuem dependências. Assim para a classificação de um determinado atributo a uma classe, é levado em consideração a probabilidade de outro atributo pertencer também à mesma classe. Já o algoritmo Naive Bayes desconsidera completamente a correlação entre as variáveis por isso recebe "naive" (ingênuo) no nome.

O classificador Naive Bayes assume que as características são condicionalmente independentes, ou seja, que a informação de um evento não é informativa sobre nenhum outro. Dessa forma, ao assumir que as características são independentes, tem-se que:

$$P(x|c_i) = \prod_{k=1}^{m} P(x_k|c_i) \tag{2}$$

sendo m o número de características dos exemplos e $P(x_k|c_i)$ a classe estimada dos exemplos de treinamento.

Apesar de sua aparente simplicidade o algoritmo Naive Bayes possui desempenho superior a outros classificadores, e uma das principais vantagens do classificador é sua capacidade de obter uma boa precisão com um pequeno número de dados de teste.

2.3 Support Vector Machine

Support Vector Machine (SVM) é um algoritmo de classificação supervisionado que, por meio de vetores de amostras de treinamento, estabelece um hiperplano ótimo de separação entre as classes a fim de maximizar a distância entre elas (HAYKIN, 2007), além de classificar adequadamente os dados de teste, os quais não foram vistos previamente.

Dado um conjunto de treinamento composto por n amostras e pertencentes a duas classes linearmente separáveis. O conjunto de treinamento é denominado por vetores de suporte e o objetivo é definir um hiperplano que separe os vetores. Entre os muitos hiperplanos possíveis, o hiperplano separador ótimo é o plano que maximiza a margem, ou seja, a distância entre o hiperplano e o vetor mais próximo de cada classe. A Figura 4 ilustra um hiperplano separador ótimo.

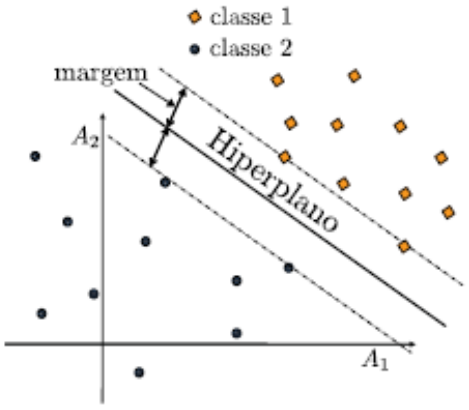

Figura 4 – Hiperplano de separação SVM de maior margem (adaptada de (HAYKIN, 2007)).

As margens grandes tendem a ter erros de generalização melhores, como pode ser explicado pelo princípio de minimização do risco estrutural (SRM). Esse princípio define um limite mais alto para o erro de generalização de um classificador a partir de seus

erros de treinamento, o número de exemplos de treinamento e sua capacidade (TAN; STEINBACH; KUMAR, 2009).

Um SVM é eficiente na classificação de dados linearmente separáveis, pela definição de um hiperplano com a maior margem. A este classificador da-se o nome de SVM linear. Também, aplica-se SVM em conjunto de dados que tenham limites de decisão não lineares, transformando os dados do seu espaço de coordenadas original x para um novo espaço Φ(x) de maior dimensão (LORENA; CARVALHO, 2007), como pode ser visto na Figura 5

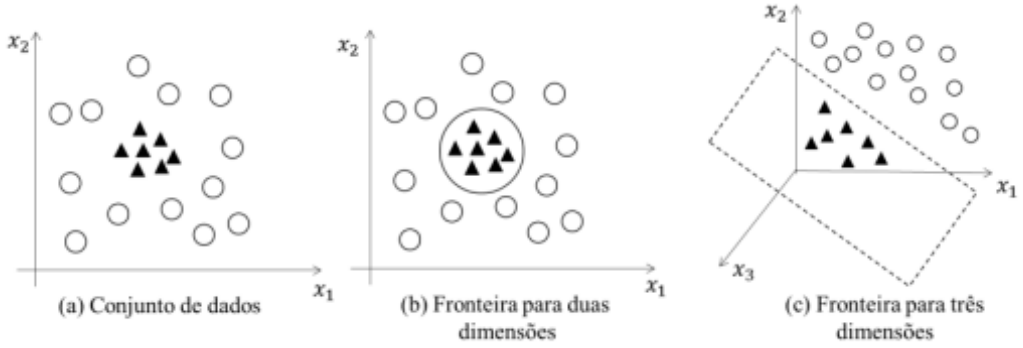

Figura 5 – Exemplo da transformação ocorrida num conjunto de dados não-linear para outro espaço de coordenadas de dimensão maior (GAMA et al., 2011).

As SVMs foram originalmente formuladas para lidar com problemas de classificação binários. Atualmente, existe uma série de técnicas que podem ser empregadas na generalização das SVMs para a resolução de problemas multiclasses (VAPNIK, 1995).

2.4 Árvores De Decisão

As árvores de decisão classificam padrões com base em uma sequência de testes e decisões. Cada caminho da raiz da árvore até uma folha corresponde a uma conjunção de testes sobre características, e a árvore como um todo corresponde a uma disjunção destas conjunções. Os padrões são classificados seguindo um caminho na árvore da raiz até uma das folhas, a qual provê a classe do padrão. Cada nó interno da árvore corresponde a um teste sobre alguma característica dos dados, e cada ramo descendente a partir de um nó corresponde a uma possibilidade de valor para a característica testada (SILVA, 2018).

A construção de uma árvore de decisão pode ser vista como um particionamento recursivo do conjunto de dados. No nó raiz todas as instâncias são consideradas e em cada nó filho considera-se somente o conjunto de dados que satisfaz a condição testada. Este processo é repetido recursivamente até alcançar um nó folha (SILVA, 2018).

Os classificadores mais conhecidos baseados em árvore de decisão são ID3, C4.5 e CART.

2.5 PCA

A Análise de Componentes Principais é a tradução do termo em inglês *Principal Component Analysis* (PCA) que é uma técnica que extrai informações importantes de um conjunto de dados multivariados, para expressá-los num conjunto de poucas e novas variáveis. Estas novas variáveis são chamadas componentes principais. Elas correspondem à uma combinação linear dos dados originais, gerando um menor número de variáveis, com perda mínima de informações.

A técnica PCA também é empregada para geração de índices e agrupamento de indivíduos. A análise agrupa os indivíduos de acordo com sua variação identificando padrões mais fortes nos dados. Ou seja, é uma técnica que agrupa os indivíduos de uma população segundo a variação de suas características (HONGYU; SANDANIELO; JUNIOR, 2016).

2.6 Métodos de Avaliação

Após desenvolver a etapa de aprendizagem de um classificador, é importante verificar se ele apresenta um bom desempenho em amostras não usadas no treinamento. Para isso, faz-se necessário definir a Matriz de Confusão responsável por gerar valores para o cálculo de importantes medidas de desempenho, além da interpretação trazida pelo gráfico da curva ROC.

2.6.1 Matriz de Confusão

Matriz de confusão é uma tabela criada para expressar visualmente o desempenho de um classificador. Em cada linha da matriz temos representada a classe das amostras que o classificador retornou e cada coluna representa a classe real. As amostras preditas corretamente recebem o rótulo de verdadeiros positivos (VP), pois o resultado do classificador respondeu a pergunta em questão com Sim (positivo) e ela é correta (verdadeira). As amostras classificadas como Sim, mas na verdade é Não, são denominadas falsos positivos (FP). De maneira similar, na segunda linha da Tabela 1 encontra amostras classificadas como Não, mas que na verdade é Sim, que são os falsos negativos (FN). Por fim, as amostras classificadas corretamente como Não são os verdadeiros negativo (VN).

Tabela 1 – Exemplo de rótulos da matriz de confusão.

	Verdadeiro	Falso
Verdadeiro	VP	FP
Falso	FN	VN

2.6.2 Medidas para Avaliação dos Resultados

Para verificar a eficiência dos algoritmos, serão calculadas algumas métricas quantitativas que serão aplicadas sob os valores resultantes do conjunto de dados de teste. As métricas são acurácia, precisão, revocação e medida F.

A acurácia (AC) quantifica a frequência com que a classificação foi realizada corretamente. Ou seja, segundo a Equação 3 a variável VP define a quantidade de verdadeiros positivos, a variável VN, a quantidade de verdadeiros negativos, e por N, a quantidade de itens do conjunto de dados.

$$AC = \frac{VP + VN}{N} \tag{3}$$

A precisão (P) calcula a porcentagem dos itens classificados como malignos ou benignos que efetivamente pertencem à estas classes. A precisão está demonstrada na Equação 4, sendo FP os falsos positivos.

$$P = \frac{VP}{VP + FP} \tag{4}$$

A revocação (R) é a proporção de verdadeiros positivos que foi classificada corretamente. A Equação 5 mostra o cálculo da revocação, contendo a variável FN que representa os falsos negativos.

$$R = \frac{VP}{VP + FN} \tag{5}$$

Já a Medida-F1, representada pela Equação 6 combina a precisão e revocação de modo a medir a qualidade geral do modelo criado.

$$F1 = \frac{2 \times P \times R}{P + R} \tag{6}$$

2.6.3 Curva ROC

A curva Receiver Operating Characteristic (ROC) é um método gráfico para avaliazação de modelos de classificação em Aprendizagem de Máquina e Mineração de Dados. A análise ROC é muito útil em problemas que existem uma grande desproporção entre as classes ou quando deve-se levar em consideração diferentes benefícios para os diferentes acertos de classificação (PRATI; BATISTA; MONARD, 2008).

O gráfico ROC é construído pela taxa de Verdadeiros Positivos (VP) no eixo das abscissas, e pela taxa dos Falsos Positivos (FP) no eixo das ordenadas. Alguns elementos no gráfico destacam-se:

- o ponto (0,0) significa que o modelo não consegue classificar nenhuma amostra como positiva;

- o ponto (0,100%) significa que o modelo classifica corretamente todas as amostras positivas e negativas;
- o ponto (100%,0) significa que o modelo sempre faz predições erradas;
- a linha diagonal entre os pontos (0,0) e (100%,100%) delimitam a área do triângulo superior da área do triângulo inferior, cujo modelo possuem pontos pertencente ao superior significa melhor desempenho que os pontos pertencentes a área inferior.

Pontos pertencentes ao mesmo espaço ROC são interpretados pela sua localização. Ou seja, quanto mais acima e à esquerda um ponto estiver do outro significa uma maior taxa de verdadeiro positivos e uma menor taxa de falsos positivos (PRATI; BATISTA; MONARD, 2008).

2.7 Bibliotecas da Linguagem Python

A linguagem de programação escolhida para desenvolver os classificadores de AM foi o Python, devido as diferentes bibliotecas de código aberto destinado ao AM. Dentre elas, destacam-se:

- scikit-learn 0.22.1
  O scikit-learn é um módulo em Python que integra uma ampla gama de algoritmos de aprendizado de máquina de ponta para problemas supervisionados e não supervisionados além de algoritmos para classificação, regressão, clusterização, redução de dimensionalidade e pré processamento. (PEDREGOSA et al., 2011).
- TensorFlow
  TensorFlow é uma biblioteca de código aberto criada para aprendizado de máquina e pesquisa de redes neurais profundas. A grande aceitação se deve muito pelo fato de seu desempenho por utilizar XLA, um poderoso compilador de álgebra linear que torna a execução mais rápida, rodando em CPUs, GPUs, TPUs e outros. O framework conta com um grande poder de abstração, permitindo que o desenvolvedor concentre na lógica de sua aplicação e não em detalhes de implementação do algoritmo. (RIBEIRO; QUIMARÃES, 2018)
- Keras
  É uma biblioteca de alto nível escrita em Python proporcionando uma interface simples para redes neurais profundas, capaz de trabalhar com bibliotecas como o Tensorflow.
- Pandas
  Pandas é uma ferramenta de manipulação e análise de dados de código aberto, rá-

pida, poderosa, flexível e fácil de usar, construída sobre a linguagem de programação Python.

- ❏ Matplotlib
  Matplotlib é uma biblioteca abrangente para criar visualizações estáticas, animadas e interativas em Python.

- ❏ Seaborn
  Seaborn é uma biblioteca de visualização de dados em Python baseada em *Matplotlib*. Fornece uma interface de alto nível para desenhar gráficos estatísticos.

2.8 Trabalhos Relacionados

Existem notáveis trabalhos que podem ser referenciados pelo emprego das técnicas deAM no reconhecimento de padrões, principalmente os estudos voltados para identificação de células cancerosas.

O trabalho de (TING; SIM, 2017) apresenta um algoritmo denominado Rede Neural Perceptron Multicamadas Autorreguladas para classificação do câncer de mama. O objetivo desse artigo foi auxiliar os médicos na identificação e classificação do câncer de mama. As imagens médicas de mama foram classificadas pela rede neural como benigna, maligna ou normal.

Os autores (SILVA; LEAL; LIMA, 2019) utilizaram modelos MLP e SVM para classificar nódulos de câncer de mama e avaliou-se o desempenho de ambos em uma base de de dados com 569 amostras. Os resultados médios obtidos no conjunto das 50 simulações realizadas, mostram que os modelos propostos apresentaram bom desempenho (todos ultrapassaram a 90,0 % de acerto) em termos da acurácia no conjunto de teste.

Os autores (MARQUES; MAGALHÃES; FERREIRA, 2019) tiveram como principal objetivo desenvolver um classificador para identificar o câncer de mama utilizando dados antropométricos e parâmetros de exame sanguíneo de rotina, que são os chamados biomarcadores. As redes neurais do tipo Perceptron Multi-Camadas e as redes Neuro-Fuzzy (ANFIS) empregados a um comitê de decisão, foram as técnicas aplicadas e mostraram como resultado uma classificação do câncer de mama, com acurácia de 97 %. Esse valor obtidos foi superior comparado aos trabalhos dos últimos anos que utilizaram biomarcadores semelhantes no período de 2013 ao início do ano de 2018.

(FREITAS; FERREIRA; SILVA, 2019) fizeram uso de redes neurais artificiais com o objetivo de reconhecimento de tumores na mama, para tal, foram utilizadas várias características, como: raio, concavidade, fractal, área, perímetro e textura. O modelo completo, em sua melhor simulação, obteve uma acurácia de 88 %, garantindo que em sua simulação mais assertiva, a chance de 88 % do modelo acertar no diagnóstico de um paciente. Em sua pior simulação, esse valor foi de 73.9 %, o que não é um valor consideravelmente baixo.

Com valores elevados para sensibilidade, em sua melhor simulação obtiveram que 96.3 % dos diagnósticos positivos para pacientes realmente doentes.

Os artigos (Bharati; Rahman; Podder, 2018), (Nematzadeh; Ibrahim; Selamat, 2015) e (Rashmi G D; Lekha; Bawane, 2015) utilizaram da base de dados gerada por Wisconsin University, cujo conjunto de dados está disponível no repositório de aprendizado de máquina da UCI. O artigo (Bharati; Rahman; Podder, 2018) utilizou dos classificados de AM Naive Bayes, Random Forest, Logistic Regression, Multilayer Perceptron, Classificadores de Vizinhos mais Próximos (*K-nearest neighbors*) na avaliação do câncer de mama. O classificador mais eficiente foi o *K-nearest neighbors* que obteve 97.9021% de acertos, seguido da rede neural MLP com 96.5035%.

Já o artigo (Nematzadeh; Ibrahim; Selamat, 2015) faz uma comparação entre os algoritmos AM usando a técnica de Validação Cruzada (*k–Fold Cross Validation (KCV)*). Os algoritmos foram Árvore de Decisão, Naive Bayes, rede neural e SVM. A análise comparativa dos estudos entre os classificadores aplicados foi focada no impacto de k na validação cruzada k-fold que obteve maior precisão.

Por fim, o artigo (Rashmi G D; Lekha; Bawane, 2015) implementou o algoritmo de classificação Naive Bayes, com o intuito de classificar e prever se as amostras da base de dados são um tumor benigno ou maligno. Os resultados apresentaram uma taxa de acerto entre 85 a 95% e uma taxa de erro entre 10 a 15%.

Segundo os trabalhos descritos anteriormente, objetivou neste implementar os algoritmos de AM: RNA, SVM, Naive Bayes e Árvore de Decisão, por serem classificadores muito estudados e empregados na classificação e predição de padrões.

Capítulo 3

Metodologia e Análise dos Resultados

Primeiramente, serão apresentadas as etapas do modelo computacional proposto, assim como as propriedades dos algoritmos de aprendizagem de máquina implementados. Na sequência, os resultados gerados serão descritos e analisados.

3.1 Metodologia

Os métodos aplicados em cada iteração do desenvolvimento deste trabalho consistem em cinco etapas, o que pode ser visto na Figura 6.

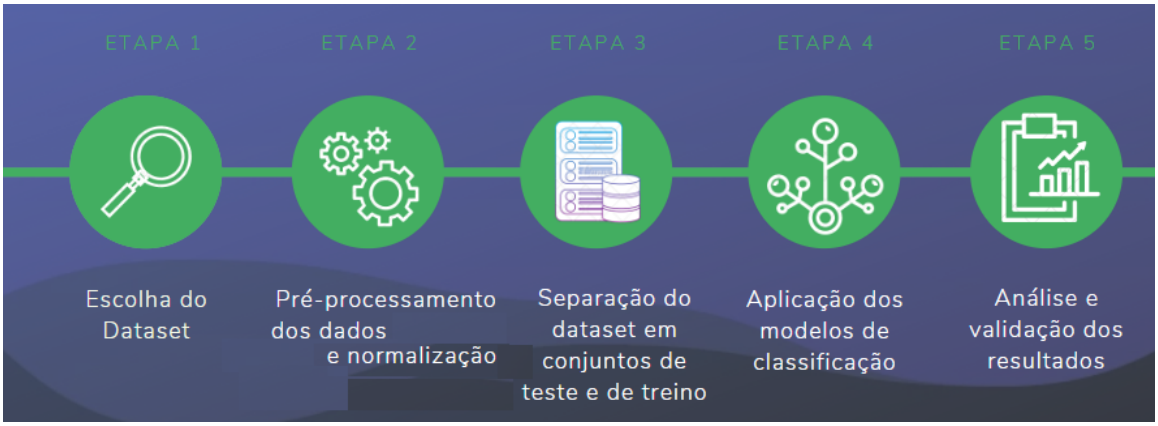

Figura 6 – Etapas da metodologia do sistema computacional proposto.

Nas próximas subseções segue a descrição de cada etapa desenvolvida pelo sistema computacional proposto.

3.1.1 Base de Dados

A base de dados escolhida foi a *Wisconsin Diagnostic Breast Cancer* uma base de dados pública disponibilizada no site UC Irvine Machine Learning Repository (WOLBERG; STREET; MANGASARIAN, (1992)). Os motivos pela escolha dela deve-se a autenticidade dos dados e sua ampla citação em trabalhos acadêmicos. Os dados são compostos

por 569 instâncias divididos entre a classe maligno, compreendendo 357 amostras, e a classe benignos, 212 amostras, conforme a Figura 7.

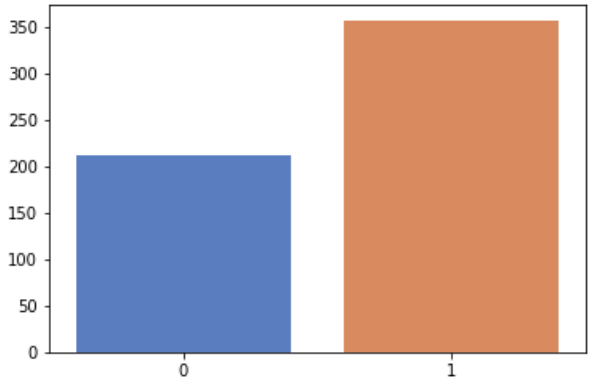

Figura 7 – Contagem das classes das instâncias (benigno ou maligno).

Os atributos que constituem a base de dados foram calculados a partir de uma imagem digitalizada de um aspirador de agulha fina sob uma massa mamária. Os atributos detalham dez características dos núcleos celulares presentes na imagem. São eles:

1. raio (média das distâncias do centro aos pontos do perímetro)
2. textura (desvio padrão dos valores da escala de cinza)
3. perímetro
4. área
5. suavidade (variação local em comprimento do raio)
6. compactação ($perímetro^2/área - 1,0$)
7. concavidade (curvatura das porções côncavas do contorno)
8. número de pontos côncavos (número de porções côncavas do contorno)
9. simetria
10. dimensão fractal da lesão

3.1.2 Pré-processamento dos dados

O primeiro passo no pré-processamento dos dados foi a substituição dos caracteres M e B pelos número inteiros 1 e 0, respectivamente. A letra M representa a classe tumor maligno e a letra B representa a classe tumor benigno.

Na sequência, o próximo passo foi a normalização dos dados para auxiliar a classificação oferecendo dados padronizados ao classificador. Por último, foi empregado a técnica

de PCA para agrupar os indivíduos segundo a variação de suas características. A Figura 8 destaca a diferença de comportamento dos atributos da base de dados sem e com o emprego do PCA.

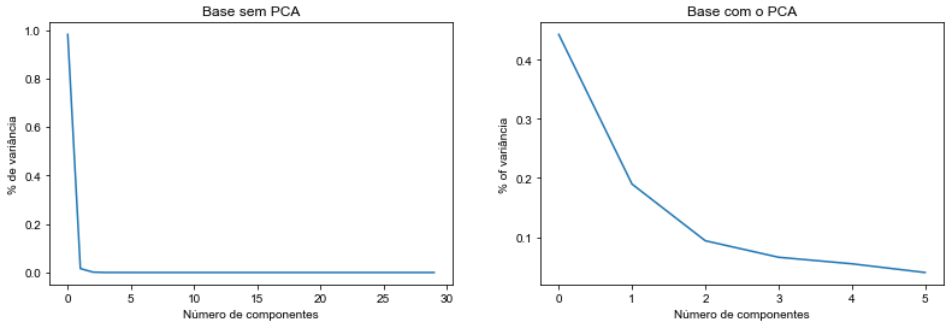

Figura 8 – Aplicação do PCA sobre as variáveis

A divisão do conjunto de treinamento e do conjunto de teste foi implementado pelo método train_test_split da biblioteca Scikit, estabelecendo uma divisão em 80% para treinamento e 20% para teste em todas as simulações de todos algoritmos.

3.1.3 Parametrização dos Classificadores de AM

Os dados de entrada para os algoritmos de Aprendizado de Máquina foram os vetores de características gerados pelo método de PCA. Assim, os parâmetros de cada classificador foram definidos por meio de tentativas e reajustes que melhor se adequaram na classificação do dados do conjunto de treinamento.

A implementação dos algoritmos dos classificadores utilizaram as funções disponíveis na biblioteca *scikit-learn* da linguagem Python.

Dentre as características relevantes de cada um dos algoritmos dos classificadores implementados foram:

- Rede Neural Artificial

  O tipo de RNA utilizado foi o MLP, cuja arquitetura possui 30 neurônios na camada de entrada, 5 neurônios em 5 camadas ocultas, 2 neurônios na camada de saída da rede, a função de ativação dos neurônios é o softmax, e o algoritmo de treinamento *Backpropagation*.

- SVM

  O modelo de treinamento utilizou a função kernel linear.

- Naive Bayes

  A estrutura do algoritmo Naive Bayes optou pela função Gaussiana para o cálculo das probabilidades.

❑ Árvore De Decisão

Esse algoritmo foi implementado utilizando a classe *DecisionTreeClassifier* contida no pacote *sklearn.tree*, que processa o algoritmo CART que constrói árvores binárias.

3.2 Avaliação dos Resultados

Os resultados gerados pelos classificadores de AM serão exibidos, primeiramente, por meio da Matriz de Confusão, como pode ser observado na Figura 9.

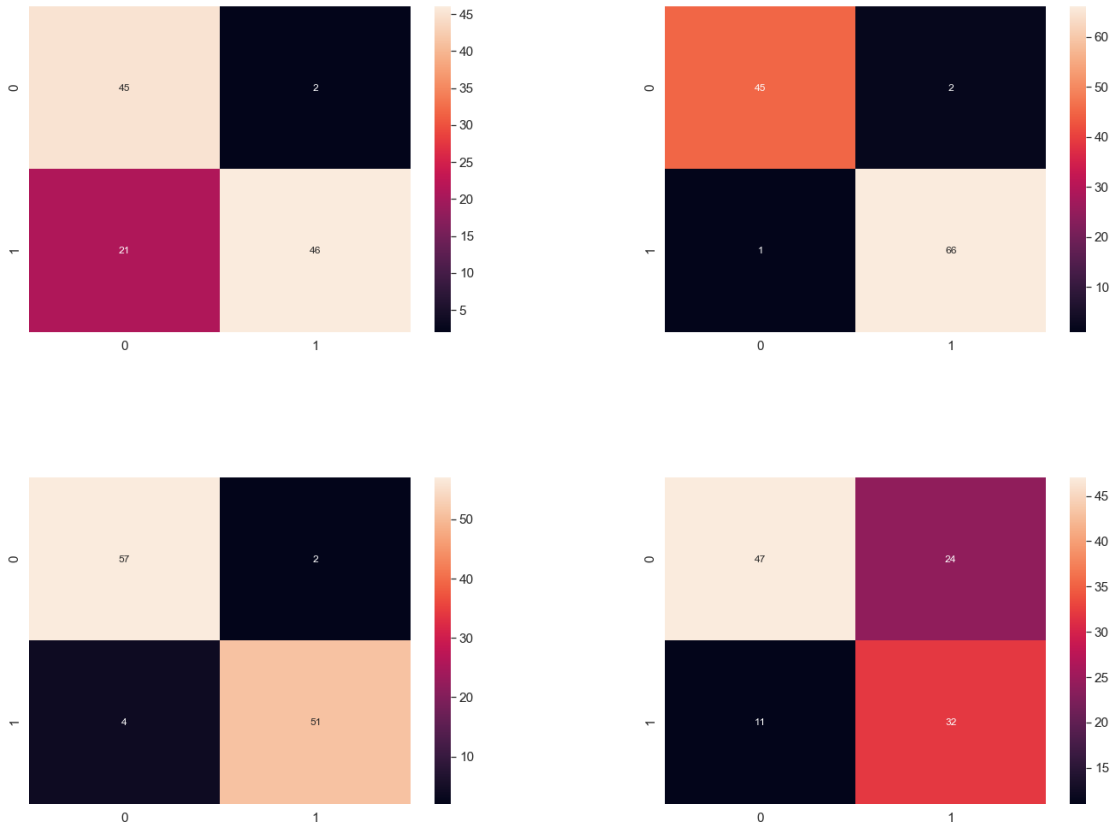

Figura 9 – Matriz de confusão gerada por cada classificador de AM.

A Figura 9 mostra o bom desempenho da RNA em classificar corretamente as classes, em comparação as demais matrizes de confusão, pois esse classificador apresentou o número elevado de verdadeiros positivos (VP) que foi de 46 e 66 para os verdadeiros negativos (VN). Ademais, o classificador RNA também evidenciou baixo número para os falsos positivos: 2 para a classe benigno e apenas 1 para a classe maligno.

As métricas de desempenho apresentadas para evidenciar a eficiência ou não dos classificadores de AM foram medidas e estão apontadas na Tabela 2.

Ao analisar a Tabela 2 observa-se que a maioria dos classificadores atingiram acurácia superior à 90%, apenas o da Árvore de Decisão que obteve 83%. Os valores alcançados

Tabela 2 – Valores das métricas de avaliação gerados pelos classificadores de AM.

Métricas	**RNA**	**SVM**	**Naive Bayes**	**Árvore de Decisão**
Acurácia	0.98	0.97	0.92	0.83
Precisão	0.97	0.93	0.95	0.75
Revocação	0.98	0.96	0.69	0.86
Medida-F1	0.97	0.97	0.79	0.83
Área ROC	0.98	0.98	0.82	0.92

pela precisão seguiram o mesmo padrão da medida acurácia, isto é, os classificadores RNA, SVM e Naive Bayes obtiveram uma porcentagem acima de 90% e o classificador da Árvore de Decisão 75%. É importante destacar que a medida precisão não avalia nenhum exemplo negativo, por isso justifica-se o desempenho insatisfatório do classificador Árvore de Decisão quando recorre-se a Matriz de Confusão da Figura 9.

Explorando os resultados da medida revocação, nota-se que os classificadores Naive Bayes e Árvore de Decisão obtiveram resultados inferiores, os quais podem ser justificados ao averiguar a Figura 9. Em outras palavras, a medida revocação inclui em seu cálculo todos os exemplos positivos, logo os classificadores RNA e SVM tiveram melhor desempenho que os outros dois classificadores.

A Medida-F1 apresentou melhores resultados para os classificadores RNA e SVM, dado que esta é uma medida decorrente da média entre a precisão e a revocação.

A Figura 10 esboça a curva ROC de todos os classificadores de AM em estudo.

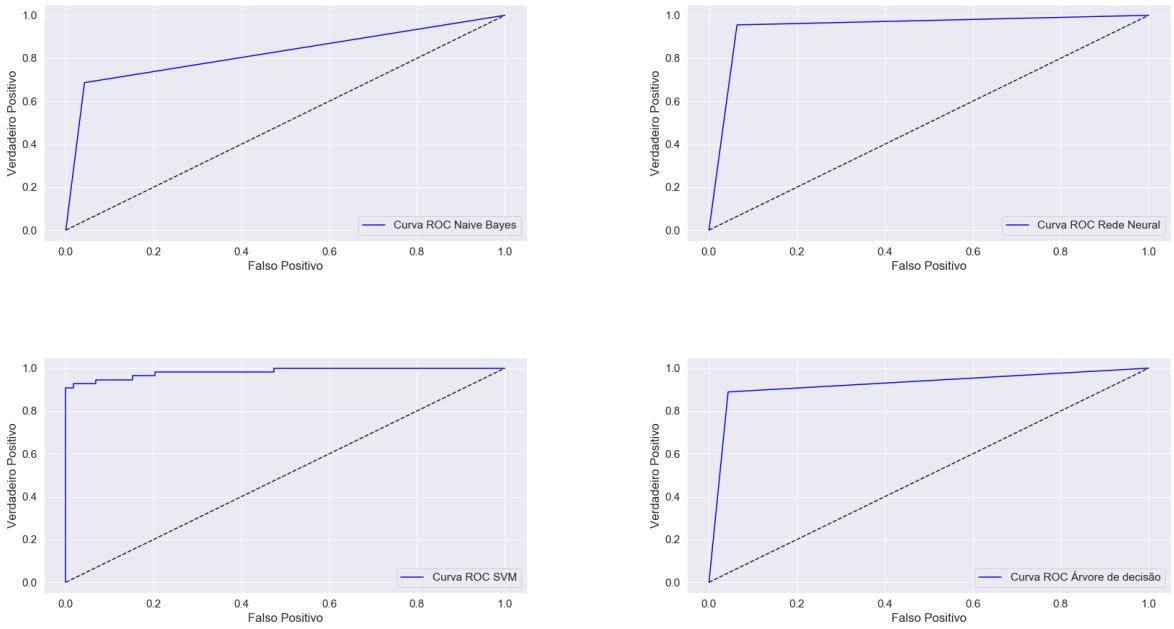

Figura 10 – Curvas ROC geradas por cada classificador de AM.

Ao comparar as curvas, conclui que os classificadores RNA e SVM foram mais eficien-

tes, pois as curvas ROC de ambos mais se aproximaram do ponto (0,10) e mais distante estiveram da diagonal principal.

3.2.1 Comparação e Discussão

A Tabela 3 expõe uma comparação de resultados gerados pelo classificador deste estudo e por outros classificadores de outros trabalhos publicados. A medida usada para comparação é a acurácia. A base de dados de todos os trabalhos da Tabela 3 foi a Wisconsin para diagnóstico de câncer de mama (WOLBERG; STREET; MANGASARIAN, (1992)).

Tabela 3 – Valores da Acurácia alcançados pelo classificador em estudo e por outros classificadores presentes na literatura para a base de dados Wisconsin.

Autor (ano)	**Método de AM**	**Acurácia %**
Freitas, Ferreira e Silva (2019)	RNA	88
Nematzadeh, Ibrahim e Selamat (2015)	RNA	98
Bharati, Rahman e Podder (2018)	RNA	96
Este estudo (2021)	**RNA**	**98**
Nematzadeh, Ibrahim e Selamat (2015)	SVM	98
Este estudo (2021)	**SVM**	**97**
Rashmi G D, Lekha e Bawane (2015)	Naive Bayes	85
Este estudo (2021)	**Naive Bayes**	**92**

De acordo com a Tabela 3, conclui que os métodos propostos neste estudo tiveram acurácia igual e superior a outros trabalhos, para algoritmos RNA, SVM e Naive Bayes.

Capítulo 4

Conclusão

Este trabalho mostra um classificador de padrões formado pelo junção do PCA e de algoritmos de AM para identificar células benignas ou malignas a partir do conjunto de características extraídas de massas celulares. Inicialmente, foi realizada a normalização dos dados. Na sequência, a técnica de componentes principais foi utilizada para agrupar e evidenciar características dos dados, e por fim a utilização dos algoritmos Árvores de Decisão, *Support Vector Machine*, Naive Bayes e Redes Neurais Artificiais para classificar as células como Maligno e Benigno.

A validação do desempenho e eficiência dos classificadores foram observadas pelas medidas acurácia, precisão, revocação, medida-F e área ROC. Os classificadores compostos pela RNA e SVM obtiveram os melhores resultados nas medidas de desempenho comparados à Árvore de Decisão e ao algoritmo Naive Bayes, o que pode ser comprovado pelas curvas e áreas desenhadas nas curvas ROC derivados destes dois. Essas conclusões devem à porcentagem acima de 92% de corretude geradas por todas as medidas de avaliação para RNA e SVM.

A técnica de PCA auxiliou o aprendizado dos algoritmos de AM, destacando sua relevância na etapa de pré-processamento do classificador. Consequentemente, agiu como facilitadora para alcançar os bons resultados.

Para trabalhos futuros, sugere-se buscar uma base de dados com maior variabilidade e trabalhar diretamente com imagens, dados reais coletados com auxílio de uma instituição de saúde para o desenvolvimento e testes sobre a base. Além disso, sugere-se definir novas arquiteturas de RNA e Algoritmos Evolutivos, tais como os Algoritmos Genéticos e o *Particle Swarm Optimization*, como classificadores de padrão.

Referências

BERTON, L. **Caracterização de classes e detecção de outliers em redes complexa**. Dissertação (Mestrado) — Instituto de Ciências Matemáticas e de Computação, University of São Paulo, São Carlos, 2011. Citado na página 15.

Bharati, S.; Rahman, M. A.; Podder, P. Breast cancer prediction applying different classification algorithm with comparative analysis using weka. In: **2018 4th International Conference on Electrical Engineering and Information Communication Technology (iCEEiCT)**. [S.l.: s.n.], 2018. p. 581–584. Citado 2 vezes nas páginas 22 e 28.

BRAY, F. et al. Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. **CA: a cancer journal for clinicians**, Wiley Online Library, v. 68, n. 6, p. 394–424, 2018. Citado na página 9.

FIDA, B. et al. Heart disease classification ensemble optimization using genetic algorithm. In: **IEEE 14th International Multitopic Conference**. [S.l.: s.n.], 2011. p. 19–24. Citado na página 10.

FREITAS, A. G. da S.; FERREIRA, P. M.; SILVA, R. M. da. Redes neurais na classificação de neoplasias mamárias. **Revista Cereus**, v. 11, n. 1, p. 140–149, 2019. Citado 2 vezes nas páginas 21 e 28.

GAMA, J. et al. **Inteligência artificial: uma abordagem de aprendizado de máquina**. Grupo Gen - LTC, 2011. ISBN 9788521618805. Disponível em: <https://books.google.com.br/books?id=4DwelAEACAAJ>. Citado 2 vezes nas páginas 5 e 17.

HAYKIN, S. **Redes neurais: princípios e prática**. [S.l.]: Bookman Editora, 2007. Citado 4 vezes nas páginas 5, 12, 13 e 16.

HONGYU, K.; SANDANIELO, V.; JUNIOR, G. AnÁlise de componentes principais: Resumo teÓrico, aplicaÇÃo e interpretaÇÃo. v. 5, p. 83–90, 07 2016. Citado na página 18.

INCA. **A situação do câncer de mama no Brasil: Síntese de dados dos sistemas de informação**. 2019. Último acesso 15 outubro 2019. Disponível em: <https://www.inca.gov.br/sites/ufu.sti.inca.local/files//media/document//a_situacao_ca_mama_brasil_2019.pdf>. Citado na página 9.

LORENA, A. C.; CARVALHO, A. de. Uma introdução às support vector machines. **Revista de Informática Teórica e Aplicada**, v. 14, n. 2, p. 43–67, 2007. ISSN 21752745. Disponível em: <https://seer.ufrgs.br/rita/article/view/rita_v14_n2_p43-67>. Citado na página 17.

MARQUES, L. S.; MAGALHÃES, R. R.; FERREIRA, D. D. Inteligência computacional aplicada à detecção de câncer de mama. **Revista Brasileira de Computação Aplicada**, v. 11, n. 1, p. 28–35, 2019. Citado na página 21.

Nematzadeh, Z.; Ibrahim, R.; Selamat, A. Comparative studies on breast cancer classifications with k-fold cross validations using machine learning techniques. In: **2015 10th Asian Control Conference (ASCC)**. [S.l.: s.n.], 2015. p. 1–6. Citado 2 vezes nas páginas 22 e 28.

OLIVEIRA, A. Izzo de. **Estudo da Relação Entre o Campo Magnético e a Intensidade de Raios Cósmicos No Meio Interplanetário Via Redes Neurais**. Tese (Doutorado), 02 2013. Citado 2 vezes nas páginas 5 e 15.

PEDREGOSA, F. et al. Scikit-learn: Machine learning in Python. **Journal of Machine Learning Research**, v. 12, p. 2825–2830, 2011. Citado na página 20.

PRATI, R. C.; BATISTA, G. E. d. A. P. A.; MONARD, M. C. Curvas roc para avaliação de classificadores. **IEEE Latin America Transactions**, 2008. Citado 2 vezes nas páginas 19 e 20.

Rashmi G D; Lekha, A.; Bawane, N. Analysis of efficiency of classification and prediction algorithms (naïve bayes) for breast cancer dataset. In: **2015 International Conference on Emerging Research in Electronics, Computer Science and Technology (ICERECT)**. [S.l.: s.n.], 2015. p. 108–113. Citado 2 vezes nas páginas 22 e 28.

RIBEIRO, M. da M.; QUIMARÃES, S. S. Redes neurais utilizando tensorflow e keras. **RE3C-Revista Eletrônica Científica de Ciência da Computação**, v. 13, n. 1, 2018. Citado na página 20.

SILVA, F. H. **Estudo e desenvolvimento de métodos para predição de doadores de sangue**. Dissertação (Dissertação de Mestrado) — Universidade Federal de Goías, 2018. Citado na página 17.

SILVA, I. N. d.; SPATTI, D. H.; FLAUZINO, R. A. **Redes neurais artificiais: para engenharia e ciências aplicadas**. [S.l.]: Artliber, 2010. Citado 3 vezes nas páginas 12, 13 e 14.

SILVA, R. M.; LEAL, M.; LIMA, F. Predico do câncer de mama com aplicação de modelos de inteligência computacional. **TEMA (São Carlos)**, SciELO Brasil, v. 20, n. 2, p. 229–240, 2019. Citado na página 21.

SOUTO, L. P. M. S. **Mineração de Imagens para a Classificação de Tumores de Mama**. Dissertação (Mestrado) — Universidade do Estado do Rio Grande do Norte, 2013. Citado na página 9.

TAN, P.-N.; STEINBACH, M.; KUMAR, V. **Introdução ao datamining: mineração de dados**. [S.l.]: Ciência Moderna, 2009. Citado 2 vezes nas páginas 13 e 17.

TING, F.; SIM, K. Self-regulated multilayer perceptron neural network for breast cancer classification. In: IEEE. **2017 International Conference on Robotics, Automation and Sciences (ICORAS)**. [S.l.], 2017. p. 1–5. Citado na página 21.

VAPNIK, V. N. **The Nature of Statistical Learning Theory**. [S.l.]: Springer-Verlag, 1995. Citado na página 17.

VIDAL, F. et al. Projeto e implementação de uma rede neural artificial para detecção do mal-posicionamento rotacional de dedos em dispositivos de captura de impressões digitais multivista sem toque. In: . [S.l.: s.n.], 2015. Citado 2 vezes nas páginas 5 e 14.

WOLBERG, W. H.; STREET, W. N.; MANGASARIAN, O. L. Breast cancer wisconsin (diagnostic) data set. **UCI Machine Learning Repository [http://archive. ics. uci. edu/ml/]**, (1992). Citado 2 vezes nas páginas 23 e 28.

Yungang Zhang et al. Highly reliable breast cancer diagnosis with cascaded ensemble classifiers. In: **The 2012 International Joint Conference on Neural Networks (IJCNN)**. [S.l.: s.n.], 2012. p. 1–8. Citado na página 10.