Proposta de Tese de Mestrado

Proposta e Avaliação de um Modelo Híbrido de Seleção de Características para o Prognóstico do Câncer de Mama

Maxwell Esdra Acioli Silva
meas@ic.ufal.br

Orientadores:

Dr. Rafael de Amorim Silva
Dr. Bruno Pimentel

Maceió
22 de Julho, 2022

Maxwell Esdra Acioli Silva

Proposta e Avaliação de um Modelo Híbrido de Seleção de Características para o Prognóstico do Câncer de Mama

Tese apresentada por Maxwell Esdra Acioli Silva em cumprimento parcial dos requisitos para o grau de Mestre em Ciências da Informática da Universidade Federal de Alagoas, Instituto de Computação.

Orientadores:

Dr. Rafael de Amorim Silva

Dr. Bruno Pimentel

Maceió
22 de Julho, 2022

Catalogação na fonte
Universidade Federal de Alagoas
Biblioteca Central
Divisão de Tratamento Técnico
Bibliotecária: Taciana Sousa dos Santos – CRB-4 – 2062

S586p Silva, Maxwell Esdra Acioli.
Proposta e avaliação de um modelo híbrido de seleção de características para o prognóstico do câncer de mama / Maxwell Esdra Acioli Silva. – 2022.
59 f. : il. color.

Orientador: Rafael de Amorim Silva.
Coorientador: Bruno Almeida Pimentel.
Dissertação (Mestrado em Informática) – Universidade Federal de Alagoas. Instituto de Computação. Maceió, 2022.

Bibliografia: f. 54-59.

1. Inteligência artificial. 2. Prognóstico. 3. Câncer de mama. 4. Aprendizagem de máquina. I. Título.

CDU: 004.8: 616.19-006

Agradecimentos

Primeiramente, a Deus que permitiu que tudo isso acontecesse, ao longo de minha vida,sem Ele nada disso seria possível. Obrigado, meu Senhor, por colocar amor, fé e esperança no meu coração.

Aos meus pais, Aleniude e Dorgival, e aos meus irmãos, Miquéias e Priscilla, por todo amor, estímulo e apoio incondicional.

À minha amada esposa Helynne, que deu forças para eu vencer essa etapa da minha vida.

Ao Profº. Dr. Rafael Amorim, pela orientação, suporte e confiança.

Resumo

A tecnologia de Inteligência Artificial tem sido fundamental no papel do cuidado à saúde da sociedade. Ela vem sendo amplamente utilizada nos diversos ramos da medicina. Uma de suas principais aplicações é no contexto do prognóstico da doença de câncer de mama. O câncer é considerado como a segunda maior causa de mortes decorrentes de doenças no mundo. Neste contexto, destaca-se o câncer de mama, que é considerado a maior ocorrência de câncer entre as mulheres no mundo. Um dos principais desafios neste cenário é identificar quais são as características mais relevantes no desenvolvimento deste tipo de neoplasia por um paciente. Este tipo de filtro é realizado pelos métodos de seleção de características. Este trabalho apresenta um modelo híbrido de seleção de características que deve ser utilizado por clínicos de um paciente, a fim de realizar uma predição de mortalidade do câncer de mama. O modelo híbrido é composto de dois algoritmos já existentes na literatura: Gain Ratio e ReliefF. Em comparação com os modelos já existentes na literatura, o algoritmo proposto apresentou melhores resultados para as métricas de avaliação utilizadas para tal finalidade.

Palavras-chave: Aprendizagem de Máquina, Prognóstico, Câncer de Mama.

Abstract

Artificial Intelligence technology has been instrumental in the role of health care for society. It has been widely used in various branches of medicine. One of its main applications is in the context of breast cancer disease prognosis. Cancer is considered the second leading cause of death from disease in the world. In this context, breast cancer stands out, which is considered the largest occurrence of cancer among women in the world. One of the main challenges in this scenario is to identify which are the most relevant characteristics in the development of this type of neoplasm by a patient. This type of filtering is performed by feature selection methods. This paper presents a hybrid feature selection model that should be used by clinicians of a patient in order to perform a breast cancer mortality prediction. The hybrid model is composed of two algorithms already existing in the literature: Gain Ratio and ReliefF. In comparison with the models already existing in the literature, the proposed algorithm presented better results for the evaluation metrics used for this purpose.

Keywords: Machine Learning, Prognosis, Breast Cancer.

Lista de Figuras

Lista de Algoritmos

Sumário

1

Introdução

No cenário de Machine Learning (ML), dados podem ser definidos como um fato, texto, imagem ou som que não foi processado. São considerados partes essenciais no contexto de ML, pois sem estes não é possível treinar o modelo, e consequentemente inferir alguma informação do objeto de estudo. Um modelo de aprendizagem de máquina é caracterizado como um algoritmo que tem a capacidade de reconhecer padrões sobre um conjunto de dados aplicados a este (Jiang et al., 2017).

Com os sucessivos avanços que houveram na Inteligência Artificial (IA) e ML, estas tecnologias passaram a ser consideradas fundamentais no papel do cuidado à saúde da sociedade, sendo amplamente utilizadas nos diversos ramos da medicina. Nesse universo, os dados do paciente, utilizados em modelos de ML, podem ser divididos em dois tipos: (i) dados clínicos e (ii) dados moleculares. Dados clínicos são aqueles coletados a partir de diagnóstico, testes laboratoriais e dados hereditários do paciente. Já os dados moleculares, também chamados de microarranjos ou dados genômicos, podem ser definidos como um conjunto de dados que contêm informações das células do indivíduo, por exemplo, a sequência de RNA (Jiang et al., 2017).

Uma das principais aplicações das tecnologias acima citadas, se dá no contexto de prognóstico e predição da neoplasia câncer. O câncer é considerado como a segunda maior causa de mortes decorrentes de doenças no mundo, de acordo com Organização Mundial de Saúde (OMS) (WHO), segundo senso realizado no ano de 2018. Neste contexto, destaca-se o câncer de mama. Este, por sua vez, é considerado a maior ocorrência de câncer entre as mulheres em todo o mundo (WCRF).

Um dos principais desafios da medicina neste cenário, é utilizar as tecnologias disponíveis para conseguir fazer o prognóstico ou predição do câncer de mama com a melhor assertividade possível. Sendo assim, já existem diversos modelos apresentados na literatura com o objetivo de ajudar neste tipo de predição. Alguns destes utilizam dados clínicos dos pacientes, outros utilizam dados moleculares dos mesmos.

Fatores prognósticos são constituídos de marcadores associados à sobrevida global, capaz de indicar como será o curso clínico da neoplasia, seja no contexto de risco de reincidência ou de morte. Estudos de fatores prognósticos permitem analisar, de forma mais detalhada, como se dará a evolução do tumor de acordo com o seu comportamento (Guerra, 2015). A classificação dos fatores prognósticos é feita com base no estado do tumor e do paciente. Fatores prognósticos relacionados ao tumor são: tamanho, tipo histológico, presença de receptores hormonais, grau de diferenciação e invasão linfonodal. Já em relação ao paciente, podem ser: histórico familiar, idade em que foi feito o diagnóstico, índice de massa corporal, além de características genéticas do mesmo (Guerra, 2015). Fatores que representam um alto risco resultando em um mau prognóstico, são: tumores maiores que 2cm de diâmetro, idade inferior a 35 anos, invasão de linfonodos axilares e invasão linfo do HER-2. Por outro lado, o prognóstico pode ser favorável quando o tumor está em estágio inicial, o tumor tem um caráter vascular, ausência de receptores hormonais e super expressão (Guerra, 2015).

Os fatores prognósticos são fundamentais para determinar qual é o melhor tipo de tratamento que deve ser aplicado em cada tipo de paciente. Possíveis tratamentos aplicados em pacientes podem ser através de realização de sessões de quimioterapia ou tratamento hormonal. Apesar destes tipos de tratamento melhorarem as chances de sobrevivência do paciente, eles podem trazer sérios efeitos colaterais à saúde da pessoa a quem estão sendo submetidos. Além de serem muito caros. Portanto, eles devem ser aplicados apenas em pacientes com um alto risco de acordo com os fatores prognósticos identificados. Porém, os fatores prognósticos tradicionais, por exemplo, estado do nódulo linfático ou tamanho do tumor, não são suficientemente precisos, necessitando assim de melhores fatores para determinar um prognóstico de alto risco (Van Diest et al., 2004). Por outro lado, estudos recentes indicam que a utilização de expressões genéticas baseadas em microarranjo podem oferecer informações prognósticas relevantes e independentes em pacientes com câncer de mama recém diagnosticadas (Kumar et al., 2012). Microarranjos é uma ferramenta laboratorial comum para detectar a expressão gênica ou mutações gênicas de alta produtividade. Estas lâminas também são conhecidas como chips gênicos ou chips de DNA. Milhares de sondas (com identidade conhecida) são imobilizadas em uma lâmina de microscópio ou chips de silício ou membrana de nylon, com milhares de pequenos pontos contendo uma sequência ou gene de DNA conhecido. Com o advento das tecnologias de sequenciamento de DNA, alguns testes para os quais eram usados microarranjos no passado agora usam tecnologias de sequenciamento em seu lugar. Mas os microarranjos são sequenciados com menor custo, portanto ainda são usados para estudos muito grandes e alguns testes clínicos (Theisen, 2008). No contexto do câncer de mama, abordagens de genoma inteiro identificaram conjuntos de genes prognósticos que preveem um intervalo curto para metástases distantes, ou seja, uma assinatura de prognóstico ruim; e descreveram perfis genéticos que medeiam metástases para um local secundário (Kumar et al., 2012).

Além disso, as células cancerígenas apresentam um fenótipo comum de crescimento celular, porém este fenótipo é capaz de surgir a partir de diferentes combinações de mutações.

Sendo assim, ao inferir como as células evoluem em tumores individuais, podemos ser capazes de identificar eventos mutacionais importantes para diferentes tipos de tumores. Podendo assim, levar potencialmente a novas terapias. (Park et al., 2008) demonstra que, é possível inferir caminhos de progressão frequentes através da utilização de microarranjos para estimar a distância entre tumores. Esta utilização pode ocorrer em sistemas de apoio à decisão (SAD). Um SAD pode ser definido como um sistema de informação que apoia as atividades educacionais, empresariais e médicas no processo de tomada de decisão. Eles unem o conhecimento tecnológico e humano para melhorar a tomada de decisão (Mazo et al., 2020). Atualmente, os SADs são utilizados em muitas áreas médicas para melhorar os serviços do sistema de saúde, o tempo de decisão e a qualidade de vida relacionada à saúde dos pacientes; tais sistemas podem reduzir tanto os custos de saúde quanto às taxas de erro médico (Mazo et al., 2020). (Mazo et al., 2020), realizou uma revisão sistemática para avaliar a disponibilidade e uso de SADs no contexto de de câncer de predições relacionadas ao câncer mama. Os trabalhos selecionados para realizar o estudo apresentam diferentes tipos de técnicas de aprendizagem de máquina, bem como diferentes tipos de dados são utilizados. Porém, um dos pontos detectados pelo trabalho em questão é que, a maioria dos SADs existentes não utilizam resultados de testes genéticos no conjunto de dados de entrada dos modelos preditivos. Possivelmente, isto deve ocorrer pelo fato deste conjunto de dados terem um número muito grande de variáveis, bem como uma quantidade bastante reduzida de amostras. Dessa forma, nota-se que seria interessante o surgimento de SADs que utilizem este tipo de informação relacionada ao tumor, para realizar inferências que ajudem os especialistas médicos no processo de tomada de decisão, quando trata-se de alternativas de tratamento em pacientes diagnosticados com câncer de mama.

Quando trata-se da utilização de dados moleculares dos pacientes, a utilização de uma ferramenta capaz de reduzir a dimensionalidade dos dados passou a ser fundamental devido ao tamanho das características contidas nestes tipos de dados. Como mencionado anteriormente, os avanços tecnológicos possibilitaram obter informações cada vez mais detalhadas acerca das células dos indivíduos, de tal forma que a quantidade de informações disponíveis chegam à casa das milhares. Portanto, é necessário utilizar algum mecanismo apto a reduzir a quantidade de informações, pois nem todas elas são relevantes para a predição da doença em questão, ou até mesmo algumas delas podem ser duplicadas. Neste cenário, surge o conceito de modelos de seleção de características, que pode ser denominada como um processamento dos dados analisados, a fim de reduzir a quantidade de características que serão utilizadas em um modelo de aprendizado de máquina, com o principal objetivo de encontrar quais são as características mais significativas dentre todas as disponíveis na análise (Visalakshi and Radha, 2014).

Segundo levantamento feito na literatura acerca de quais são os métodos de seleção de características mais utilizados neste tipo de prognóstico, destacam-se os métodos de seleção de características: (i) Relief-F; (ii) Information Gain; e (iii) Gain Ratio (Silva et al., 2021).

Portanto, este trabalho apresenta um método híbrido de seleção de características, utili-

zando dados clínicos e moleculares dos pacientes que pode ser utilizado para fazer predições do estado clínico de um paciente diagnosticado com câncer de mama. O cenário utilizado neste trabalho, está relacionado à realização da predição do risco de morte de um paciente, uma vez que este apresenta a neoplasia em questão. O algoritmo híbrido apresentado é composto de dois algoritmos já existentes na literatura: Relief-F e Gain Ratio.

A proposta deste trabalho, visa diminuir o impacto dos principais problemas de cada um dos algoritmos, quando utilizados de forma isolada. Pois o modelo de seleção *Gain Ratio* é a modificação do algoritmo de *Information Gain*, que tem o objetivo de reduzir o viés causado por este último. Porém, uma de suas grandes desvantagens é que ele calcula um peso para um recurso sem examinar outros recursos disponíveis. Se os recursos forem dependentes, isso geralmente não será refletido em seus pesos. Um recurso que contém algumas informações sobre a classe de classificação por conta própria, mas nenhum quando outro recurso mais informativo estiver presente, receberá um peso diferente de zero. Os recursos que contêm pouca informação sobre a classe de classificação receberão um peso pequeno, mas um grande número deles ainda pode anular recursos mais importantes. Esses dois problemas terão uma influência negativa na precisão da classificação, principalmente quando houver muitos recursos disponíveis (Karegowda et al., 2010b). Já o Relief-F é utilizado puramente como método de seleção de características, traz consigo o problema de não fazer uma boa avaliação qualitativa de características que têm uma certa dificuldade de distinção, ou seja, causando uma ineficiência de avaliação de classes minoritárias (Urbanowicz et al., 2018b).

A metodologia utilizada no experimento de trabalho consiste em definir o fluxo de funcionamento do algoritmo proposto, estabelecer quais serão os modelos preditivos utilizados para obter os resultados de predição sobre o conjunto de dados resultantes das três abordagens de seleção de características discutidas neste trabalho. Através dela, também deverá ser estabelecido qual deve ser a densidade da quantidade de características que vai trazer os melhores resultados possíveis para o prognóstico de câncer de mama, em específico a predição da mortalidade em decorrência do câncer através da utilização de dados genômicos dos pacientes.

Além disso, o algoritmo apresentado tenta diminuir o impacto dos pontos negativos de cada modelo em separado, criando um novo modelo que é capaz de levar em consideração as características que têm dependência entre si, bem como apresentar boas avaliações qualitativas das características mesmo em *datasets* como um número muito elevado de características, que o caso de *datasets* contendo informações genéticas dos pacientes. Além dos ganhos mencionados anteriormente, o modelo proposto é possível trazer ganhos consideráveis na questão do processamento do modelo, acurácia e interpretabilidade dos dados que estão sendo utilizados no SAD. Pois, quando trata-se de dados moleculares, que têm um valor alto de dimensão, com a redução da quantidade de genes, o modelo necessitará processar menos informações, resultando em um ganho de processamento. Também é possível obter o ganho de acurácia, em relação aos modelos utilizados de forma isolada, como será apresentado na seção de resultados. Outro fator importante, é a questão da interpretabilidade dos dados que estão sendo

analisados no contexto de prognóstico, pois como haverá redução dos genes, resultando nos mais importantes, no contexto da predição a ser realizada. O especialista de saúde poderá obter uma interpretação mais assertiva sobre quais fatores são mais relevantes para a predição que deseja ser feita.

O modelo proposto tem o propósito de ser utilizado por cientistas de dados que, ao implementar um sistema de apoio à decisão médica, poderão utilizar o modelo de aprendizagem apresentado na construção do SAD. O algoritmo poderá ser implementado através de bibliotecas renomadas de aprendizagem de máquina, tais como Sklearn (Pedregosa et al., 2011) ou Orange (Demšar et al., 2013), em ambos os casos o ambiente de desenvolvimento é voltado para a linguagem de programação Python.

1.1 Proposta

A proposta deste trabalho consiste em fazer uma revisão sistemática da literatura e investigar quais são os algoritmos mais utilizados no contexto de prognóstico de câncer de mama, a partir dos resultados da pesquisa realizada, avaliar quais são pontos positivos e negativos para os modelos encontrados. Após isso, este trabalho visa definir aspectos que possam preencher possíveis lacunas existentes na utilização dos modelos já existentes na literatura, uma solução possível seria através da definição de um novo algoritmo que solucione aspectos negativos dos achados científicos na literatura. Após a conclusão da revisão sistemática de literatura, foram encontrados os algoritmos mais usados no contexto de prognóstico de câncer de mama, são eles: Gain Ratio e Relief-F. Uma vez definidos os objetos de estudos, foram encontrados os pontos fortes e fracos de cada um destes algoritmos quando aplicados de forma isolada na seleção de características de um conjunto de dados pré definido. Portanto, este trabalho tem como proposta principal a definição de um novo modelo, este por sua vez como sendo um modelo híbrido, ou seja, a junção dos dois modelos já existentes na literatura. Ainda neste cenário, este trabalho tem o objetivo de verificar se a aplicação deste tipo de abordagem pode gerar resultados mais satisfatórios no contexto de predição sobre a doença de câncer de mama. Espera-se que tais resultados se demonstrem através de melhores valores para as métricas de avaliação de modelos de aprendizagem de máquina quando aplicado o modelo de seleção de característica proposto neste referido trabalho. Também deve fazer parte do objeto deste trabalho a identificação das limitações da aplicação do modelo proposto, bem como estabelecimento de quais são os pontos que devem ser abordados futuramente, como forma de preenchimento de eventuais pontos limitantes da pesquisa realizada neste estudo.

1.2 Hipóteses

1.2.1 Hipótese nula (Ho)

A técnica de seleção de características obtida através da composição das técnicas já existentes Relief-F e Gain Ratio, não apresenta um resultado superior às abordagens já existentes na literatura no contexto de prognóstico de câncer de mama.

1.2.2 Hipótese alternativa (Ha)

A composição das técnicas de seleção de características Relief-F e Gain Ratio, quando aplicadas no contexto do prognóstico do câncer de mama, apresenta uma eficiência maior quando comparada às abordagens de seleção de características atualmente existentes.

1.3 Objetivo Geral

O nosso trabalho é voltado para uma solução que visa o aumento da interpretabilidade do modelo utilizado para realizar o prognóstico da doença de câncer de mama, através da redução de dimensionalidade dos dados trabalhos por este. As atividades da nossa proposta podem ser descritas da seguinte maneira:

1.4 Objetivos Específicos

- Revisão sistemática da literatura para identificar quais são as técnicas mais utilizadas neste contexto;
- Definição de um novo modelo híbrido utilizando as técnicas de seleção de características já existentes Relief-F e Gain Ration;
- Obtenção de um conjunto de dados contendo informações relevantes para predição da recorrência do câncer de mama;
- Adaptação dos dois modelos de selecionados para construção do modelo híbrido proposto;
- Teste do modelo híbrido proposto;
- Extração de características mais relevantes no prognóstico da recorrência do câncer de mama;
- Comparação dos resultados obtidos utilizando o método híbrido com os métodos existentes na literatura.

1.5 Estrutura do Trabalho

Este trabalho está estruturado da seguinte forma: o capítulo 2 detalha a revisão sistemática feita para obtenção dos principais trabalhos desenvolvidos no contexto do prognóstico da neoplasia de câncer, também temos o capítulo 3 que apresenta elementos importantes que são bastantes relevantes no entendimento teórico, tais como fatores prognósticos para a doença de câncer de mama. Além de trazer uma breve explicação sobre algoritmos de seleção de características. Já o capítulo 4 traz uma breve visão geral sobre os algoritmos utilizados para a construção do algoritmo proposto neste. No capítulo 5 apresentamos o conjunto de dados utilizados no experimento, bem como os resultados obtidos e uma discussão sobre os mesmos. Por fim, o capítulo 6 traz as considerações finais da proposta.

2

Trabalhos Relacionados

A utilização de técnicas de seleção de características na detecção de propriedades determinantes no prognóstico do câncer vêm sendo estudadas ao longo dos anos, muitos trabalhos sobre este tema foram publicados. Neste contexto, foi realizada uma revisão sistemática de literatura, com o objetivo de identificar trabalhos que tratam da utilização de técnicas de seleção de características no contexto do prognóstico da doença de câncer (Silva et al., 2021). Uma revisão sistemática é uma revisão da literatura realizada a partir de uma pergunta de pesquisa definida, por meio da qual se busca identificar, avaliar, selecionar e sintetizar evidências de estudos empíricos que atendam a critérios de elegibilidade predefinidos (Garcia, 2014).

A revisão sistemática realizada considerou os 27 itens da recomendação PRISMA (Moher et al., 2011) e se baseou no protocolo definido por Kitchenham e Charters (Kitchenham and Charters, 2007). Além disso, foram utilizadas ferramentas como: a) Mendeley, para realizar a organização dos artigos científicos; b) Google Sheets, usada para organizar e sintetizar os achados científicos. A recomendação PRISMA consiste em um checklist contendo 27 itens e tem o objetivo de ajudar os autores a melhorarem o relato de revisões sistemáticas e meta-análises. O foco foi em ensaios clínicos randomizados, mas o PRISMA também pode ser usado como uma base para relatos de revisões sistemáticas de outros tipos de pesquisa, particularmente avaliações de intervenções. O PRISMA também pode ser útil para a avaliação crítica de revisões sistemáticas publicadas. Entretanto, o checklist PRISMA não é um instrumento de avaliação de qualidade para ponderar a qualidade de uma revisão sistemática (Galvão et al., 2015).

O protocolo utilizado na revisão sistemática em questão, consiste dos seguintes elementos: (i) elaboração das questões de pesquisa; (ii) definição das palavras chaves; (iii) escolha das fontes científicas; (iv) definição dos critérios de inclusão e exclusão dos artigos relacionados nesta revisão; e (v) estratégia utilizada na busca dos artigos investigados neste trabalho. Vale ressaltar que, toda revisão sistemática deve-se basear em questionamentos que norteiam a busca por informação nos artigos investigados para responder apropriadamente a cada questão

levantada. Neste contexto, a figura a seguir identifica as principais etapas do protocolo realizado nesta revisão sistemática (Silva et al., 2021).

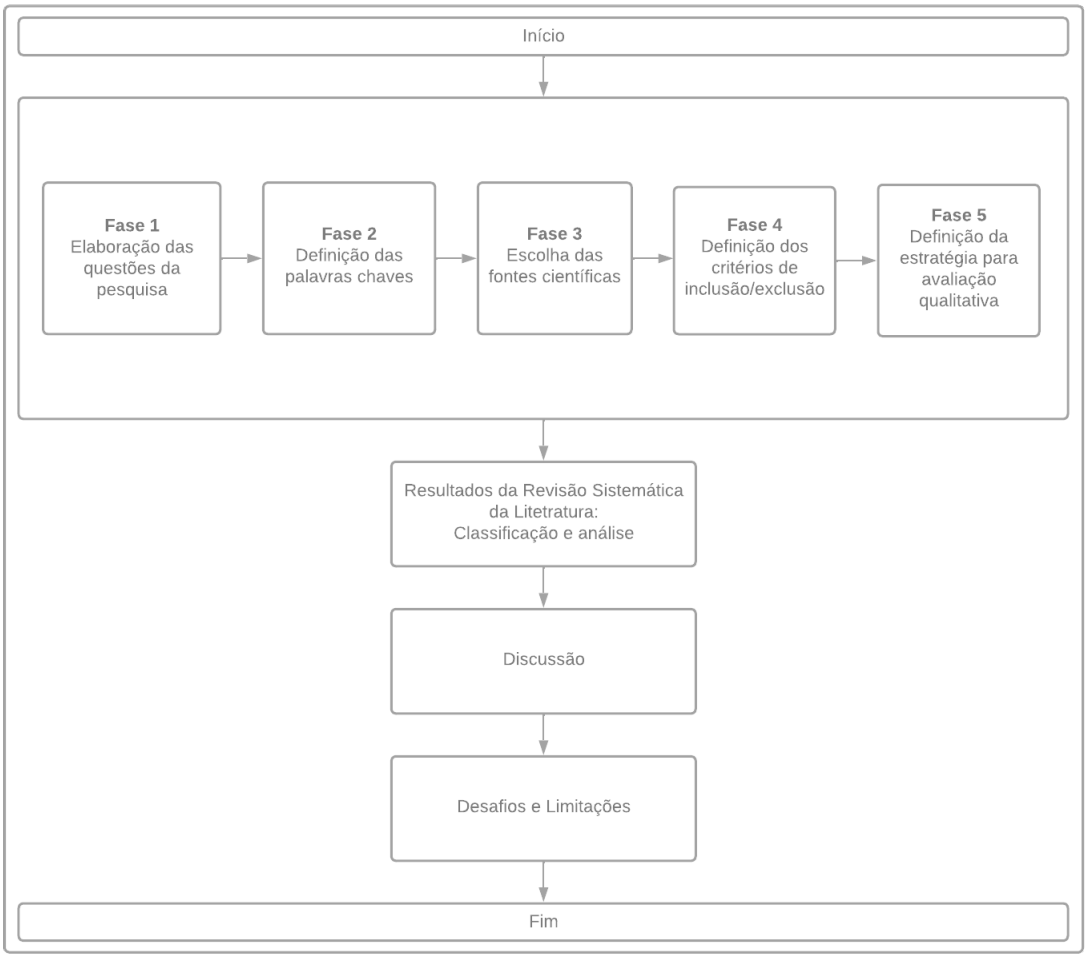

Figura 2.1: Diagrama com o fluxo do protocolo utilizado na revisão sistemática que serviu como base para este projeto

Para definição de um nível de qualidade sobre os artigos investigados na revisão realizada, foram definidos três critérios de elegibilidade, a saber: (a) critérios de inclusão; (b) critérios de exclusão; e (ci) critérios de qualidade. Onde nos critérios de inclusão, foram considerados apenas trabalhos científicos primários publicados em periódicos ou anais de eventos e escritos na língua inglesa. Já nos critérios de exclusão, foram eliminados trabalhos que não atendiam aos critérios de inclusão. Sobre os critérios de elegibilidade, foi definido um esquema de pontuação, que considera a estratégia PICOS. Estabelecendo assim, um limiar sobre a pontuação atingida por cada artigo selecionado. A respeito da estratégia PICOS, que é um acrônimo para Paciente (Pacient, em inglês), Intervenção (Intervention), Comparação (Comparison), Resultado (Outcome) e Assunto (Subject). Estes cinco componentes são os elementos essenciais da questão de pesquisa para busca bibliográfica de evidências (Santos et al., 2007). Neste cenário, cada critério PICOS é analisado e uma nota de 0 a 1 (i.e. 1 se o critério for atendido, 0.5 se for

parcialmente atendido e 0 se o critério não for atendido) é dada para cada critério de qualidade, elaborando uma média de qualidade. Os trabalhos que não obtiveram uma média de qualidade mínima de 3 pontos foi excluído do processo de seleção.

As fontes de informação utilizadas na busca dos trabalhos investigados foram feitas em plataformas como: (i) ACM; (ii) PubMed; (iii) IEEE; (iv) Springer; e (v) Elsevier (através do Science Direct). Além disso, o protocolo de busca foi executado no dia 25 de setembro de 2020. Para realizar esta revisão sistemática, utilizou-se apenas bases eletrônicas em motores de busca da World Wide Web. A string de busca foi elaborada considerando as palavras chaves retiradas das questões de pesquisa. A seguinte string de busca foi utilizada: ("machine learning") AND (("feature selection") OR ("features selection")) AND ((method) OR (algorithm) OR (technique)) AND (("cancer prognosis") OR ("cancer prediction")).

O protocolo foi executado de acordo com as configurações apresentadas acima e a seleção de estudos ocorreu em 6 etapas. Na etapa 1, a string de busca é aplicada nos motores de busca das fontes de informação, coletando 301 artigos destas fontes. Na etapa 2, utilizou-se os filtros existentes nestas fontes de pesquisa para coletar apenas os estudos publicados em anais e periódicos e definir o intervalo de tempo de publicação de acordo com o critério de inclusão. Assim, estudos não revisados por pares como enciclopédias, normas, cursos, resenhas, livros, entre outros, foram excluídos no processo. A etapa 3 corresponde a exclusão de todos os artigos duplicados, redundantes e indisponíveis. Na etapa 4, realizou-se uma revisão dos títulos, das palavras-chave, do local de publicação e do resumo para excluir os artigos que não atendessem aos critérios de inclusão e exclusão. Na etapa 5, todos os artigos incluídos na etapa 4 foram recuperados de suas respectivas fontes de informação e lidos. Na etapa 6, foi feita uma avaliação da qualidade dos artigos considerando a inclusão de elementos relacionados à abordagem PICOS. Uma média foi calculada para cada artigo e foram selecionados apenas os que atingiram a pontuação mínima de 3 pontos (Silva et al., 2021).

O processo de obtenção sistemática de trabalhos que correspondem às questões de pesquisa obteve os seguintes resultados: na etapa 1, um total de 628 artigos foram coletados pelas fontes de informação. Em seguida (na etapa 2), 301 artigos foram incluídos após a aplicação dos filtros, sendo 327 artigos eliminados por não serem artigos primários, ou por não pertencerem a periódicos ou conferências. Na etapa 3, foram removidos os artigos cujo título não estava de acordo com o trabalho proposto, neste caso foram excluídos 216 artigos. Já na etapa 4, foram removidos 2 artigos duplicados. Na etapa 5, 24 artigos foram incluídos após análise dos seus respectivos resumos. Por fim, na etapa 6, foram removidos 3 artigos completos cujo objetivo não estava de acordo com a proposta deste trabalho. Ao final das etapas anteriores, restaram 21 artigos completos para serem analisados qualitativamente de acordo com os critérios do análise qualitativa de todos os artigos incluídos de acordos com as etapas anteriores da estratégia PICOS, porém nenhum artigo foi excluído por não atingir a nota mínima definida para estes critérios, ou seja, a nota mínima de 3 pontos (Silva et al., 2021).

Através da revisão foi possível identificar que o principal tipo de dados utilizados no prognós-

tico de câncer são dados moleculares da doença. Um dado molecular pode ser definido como um dado obtido a partir da utilização de técnicas de biotecnologia a fim de extrair informações mais detalhadas acerca da doença analisada, por exemplo, uma sequência de DNA ou um conjunto de microarranjos das células. Também foi identificado que as técnicas mais utilizadas são: Relief-F, Information Gain, Gain Ratio, Random Forest, T-Test e Fast Correlation- Based Feature (FCBF) (Silva et al., 2021).

Um dos principais trabalhos analisados na revisão, que também está inserido no contexto de câncer de mama, é o trabalho desenvolvido por (Khourdifi and Bahaj, 2018), ao qual realiza uma pesquisa focada no uso do FCBF para para remover características irrelevantes e redundantes que ajudarão a melhorar a precisão preditiva dos classificadores, e então usamos técnicas de classificação em ciência médica e bioinformática, no contexto de câncer de mama. Onde primeiramente é feita uma classificação do conjunto de dados e depois determina o melhor algoritmo para o diagnóstico e a previsão da doença do câncer de mama. A previsão começa com a identificação de sintomas em pacientes, identificando em seguida pacientes doentes de um grande número de pacientes doentes e saudáveis. O mesmo utiliza a base de dados disponíveis publicamente na Universidade de Wisconsin Hospitais Madison a respeito de câncer de mama (Lichman et al., 2013). A base de dados é composta 699 amostras, contendo informações como médias da simetria, concavidade, perímetro, raio dos nódulos detectados no paciente. Os dados também incluem informação acerca da natureza da neoplasia, ou seja, se a mesma é benigna ou maligna. Uma vez feito o pré-processamento da base de dados, é feito o processo de seleção de características através do modelo FCBF, onde o mesmo aplicou a validação cruzada k-fold, com k igual a 10, que é uma técnica utilizada para avaliar modelos preditivos que dividem o conjunto original em um treinamento amostra para formar o modelo, e um conjunto de testes para avaliá-lo. Depois de aplicar os métodos de pré-processamento e preparação, foi feita uma análise visual dos dados, a fim de determinar a distribuição de valores em termos de eficácia e eficiência. Avaliando-se os valores de eficácia de todos os classificadores em termos de tempo para construir o modelo, instâncias corretamente classificadas, classificadas incorretamente instâncias e precisão (Khourdifi and Bahaj, 2018). O mesmo também utilizou outros modelos para fazer a comparação entre o método proposto e os já existentes na literatura, os modelos utilizados foram: K-NN, SVM, Random Forest, Naive Bayes e MLP. Após a realização da comparação entre o modelo proposto e os demais, o método FCBF se mostrou eficaz e relevante na eliminação de elementos redundantes e irrelevantes acerca de características utilizadas para fazer predição no contexto da doença de câncer de mama.

3

Background

Segundo a Organização Mundial de Saúde (OMS), de acordo com senso realizado no ano de 2020, o câncer é considerada com a segunda maior causa de mortes decorrentes de doença no mundo (WHO). Diante deste cenário, o câncer de mama, por sua vez, é considerado como o de maior ocorrência entre as mulheres em todo o mundo. Estima-se a ocorrência de aproximadamente 9,9 milhões de mortes no mundo em decorrência da doença de câncer, dentre este total de mortes, cerca de 6,9% destas mortes são de câncer de mama (Sung et al.).

3.1 Câncer de mama

Durante o decorrer do último século o conhecimento médico acerca da doença de câncer evoluiu consideravelmente, com esta evolução foi possível obter uma maior compreensão sobre o mesmo. Isso foi possível graças ao surgimento de novas tecnologias terapêuticas e diagnósticas, as quais possibilitaram descobrir de uma forma mais antecipada a existência de células cancerígenas nos pacientes, bem como, fornecer qual a terapia específica para o tipo de câncer em questão. No Brasil, as primeiras preocupações acerca da doença de câncer surgiram em meados de 1920. Porém, só a partir de 1940 é que surgiram as primeiras instituições especializadas no estudo acerca da doença de câncer. Além disso, foi a partir deste período que, iniciaram-se as primeiras campanhas educativas, destacando a importância do diagnóstico como a forma mais efetiva no tratamento da doença, pois quanto mais cedo descobrir a doença, maior são as chances de cura (Teixeira and AraÃ Neto, 2020).

A possibilidade de obtenção do diagnóstico do câncer de mama no Brasil, ganhou outro fator importante a partir da inclusão de imagens de exames que possibilitavam a visualização das primeiras lesões mamárias. A partir daí houve uma maior mobilização para a atenção acerca da saúde da mulher, consequentemente houve um maior rastreamento do câncer de mama (Teixeira and AraÃ Neto, 2020). O câncer de mama é causado pela multiplicação desordenada das células mamárias, gerando células anormais que se multiplicam, gerando um tumor (INCA).

Existem diversos tipos de exames que permitem o diagnóstico de nódulos mamários. Os principais são: a mamografia, ultrassonografia e ressonância magnética. A mamografia é considerada a mais importante entre os exames, pois trata-se do método mais indicado na avaliação de alterações na mama de pacientes assintomáticas. Este tipo de técnica, atualmente possibilita dois tipos de formação de imagens, a primeira é formada pelo conjunto *filme-écran*, que é considerada a mamografia convencional. Já o segundo tipo de imagem é obtido a partir de um receptor digital, esta também conhecida como mamografia digital (Chala and Barros, 2007).

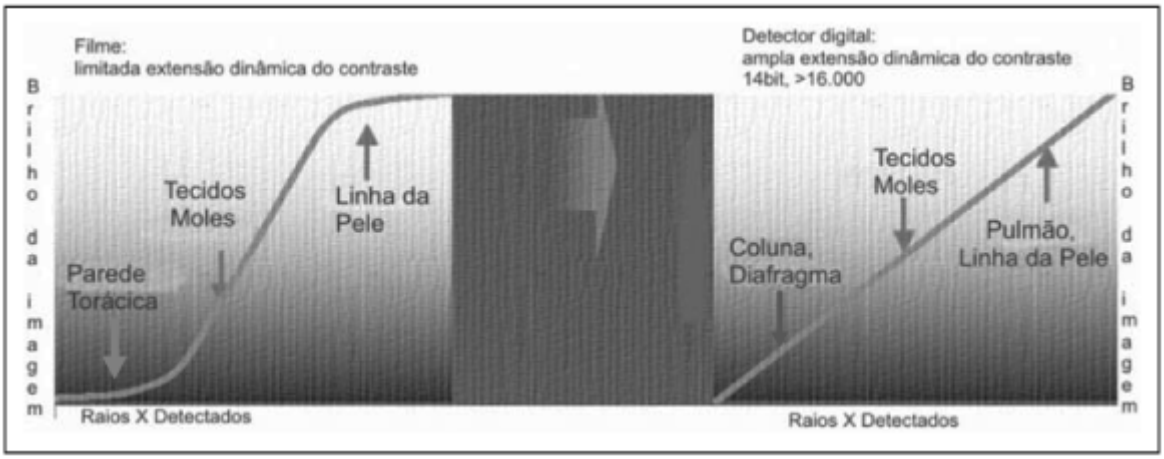

Figura 3.1: Comparação da extensão dinâmica do contraste na aquisição digital e de filme da imagem da mama

(Freitas et al., 2006)

Na mamografia convencional, o meio de aquisição da imagem, de exposição e armazenamento da mesma é representado pelo filme, este por sua vez não deixa uma margem para melhoria da imagem gerada, apesar de fornecer um bom contraste e uma boa resolução espacial. Já na mamografia digital, a aquisição, exposição e armazenamento das imagens são feitos em processos distintos. O que possibilita um melhor aperfeiçoamento das imagens, além disso o processo de análise da imagem gerada é feita através da utilização de um monitor de alta resolução, o que permite melhorar o contraste das imagens, a Figura 3.1 demonstra os níveis de contraste nos dois tipos de imagens. Porém a capacidade de detecção do câncer de mama através de imagens de mamografia varia de paciente para paciente, o mais importante destes fatores é a densidade radiológica da mama (Chala and Barros, 2007). A Figura 3.2 demonstra a comparação entre imagem da mama obtida a partir de uma mamografia convencional e outra a partir do método digital.

O principal método adjacente a mamografia na detecção do câncer de mama é a ultrassonografia. Que é um método que obtém uma boa imagem dos tecidos mamários. Este tipo de método é feito através de um aparelho que emite ondas sonoras de alta frequência. A vibração dos tecidos produz um eco, que por sua vez é lido pelo aparelho e consequentemente convertido em imagem. Este tipo de imagem é mais indicado quando o objetivo é diferenciar e caracterizar nódulos sólidos e cistos previamente identificados pela mamografia. Por conta

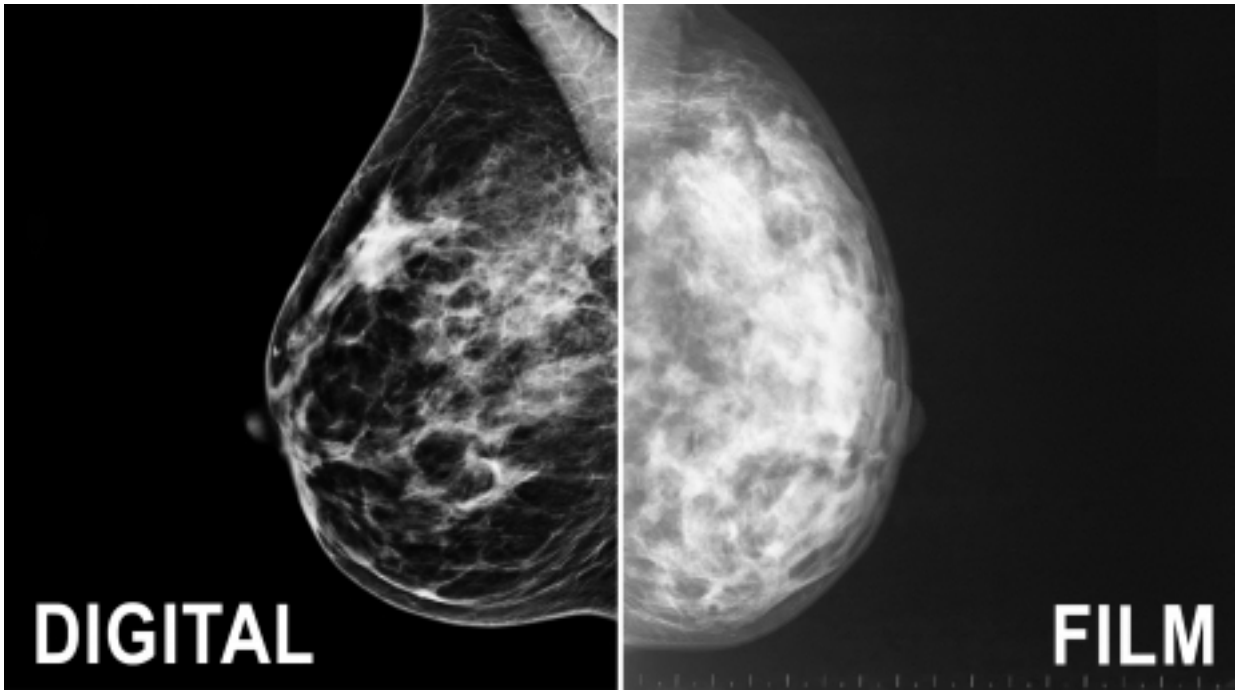

Figura 3.2: Comparação entre imagens de mamografia convencional e digital (BedfordBreastCancer)

disto, esta técnica é tida como um método suplementar à mamografia. A Figura 3.3 demonstra aspecto característico de cisto simples à esquerda (seta branca), que é bem circunscrito com margem imperceptível, anecoica, e demonstra por transmissão. Isso está em contraste com o câncer de mama à direita (seta branca), que é mal margeado, de forma irregular, hipoecoico e não apresenta transmissão por via.

Por fim, a ressonância magnética é outra forma diagnosticar o câncer de mama. Este tipo de exame também é considerado uma forma auxiliar no diagnóstico da doença acompanhado de uma mamografia e a ultrassonografia. Nele é utilizado um aparelho de ressonância magnética, que faz o uso de ímãs para gerar as imagens. Ele vem sendo bastante utilizado em pacientes com um alto risco de desenvolvimento desta neoplasia (Chala and Barros, 2007). Na Figura 3.4 podemos verificar algumas imagens de uma ressonância magnética da mama.

Diante disto, verifica-se que é necessário aplicar o tipo de exame mais apropriado para o paciente em questão, para fazer o diagnóstico precoce do câncer de mama. Uma vez que foi feito o diagnóstico, é possível aplicar as medidas terapêuticas mais assertivas para o caso da paciente, bem como utilizar alguma técnica disponível a fim de obter o quadro prognóstico da doença.

3.2 Fatores prognóstico para o câncer de mama

O curso clínico da doença de câncer de mama, bem como sua sobrevida, podem variar de paciente para paciente, segundo a história natural da neoplasia. Tal variação pode ser determinada

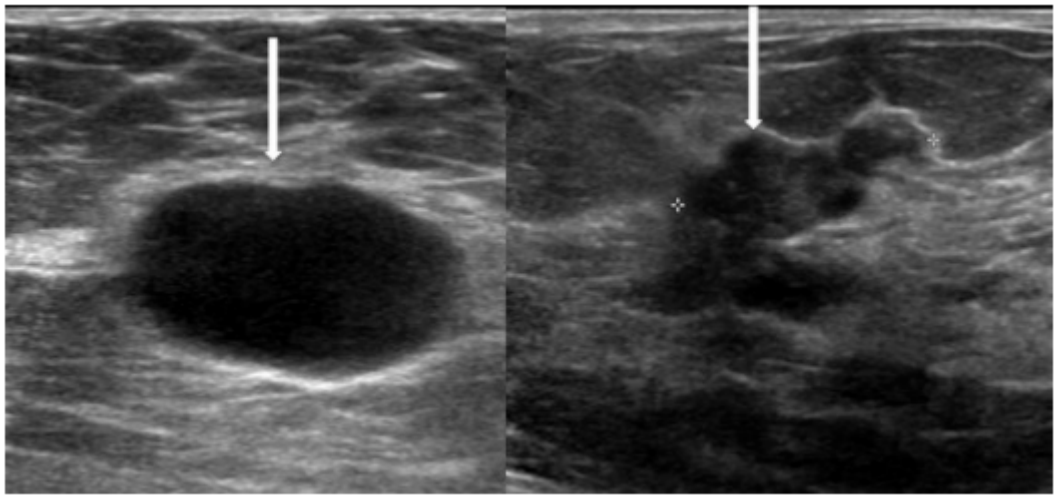

Figura 3.3: Demonstração de um cisto identificado na mama através de uma imagem de ultrassom

(Dunne et al., 2017)

por uma séria de fatores complexos, por exemplo, a diferença na rapidez da duplicação tumoral, a capacidade de metastização do nódulo, ou outros fatores que, por ora, ainda não são totalmente compreendidos, que estão relacionados à condição hormonal, nutricional e imunológica do paciente. Porém, alguns fatores anatômicos continuam sendo de fundamental importância na avaliação prognóstica da doença, tal como o tamanho do nódulo primário e a condição dos linfonodos. Além disso, fatores relacionados às características biológicas e histológicas do tumor também são determinantes para o prognóstico da evolução do câncer (Freitas JÃet al., 2017).

Os fatores prognósticos do câncer de mama podem ser úteis em três situações segundo Clark (Clark et al., 1994): a primeira diz respeito a identificação de pacientes onde o prognóstico é suficiente, e nenhum tratamento adjuntivo ao tratamento cirúrgico agregará algum benefício ao paciente; a segunda diz respeito a obtenção de um prognóstico ruim em relação ao tratamento convencional, que qualquer outra forma de tratamento mais intensa deveria ser aplicado; a terceira, é aquela cujo o prognóstico é capaz de indicar uma terapia específica para um paciente específico. Um fator prognóstico pode ser definido como um parâmetro possível de ser mensurado no momento do diagnóstico e que serviria como preditor da sobrevida ou do tempo livre de doença.

O objetivo dos cuidados à saúde é prevenir, tratar e entender doenças humanas. Neste cenário, um médico através da experiência e educação é capaz de desenvolver, um entendimento de como aplicar procedimentos e métodos de diagnóstico e prognóstico, bem como se dará o efeito dos medicamentos e tratamentos em pacientes. No processo de tomada de decisão, é exigido do médico a capacidade de analisar e compreender o estado clínico do paciente, bem como entender qual a necessidade da sua intervenção e quais ações ele pode intervir (Bronzino, 2000).

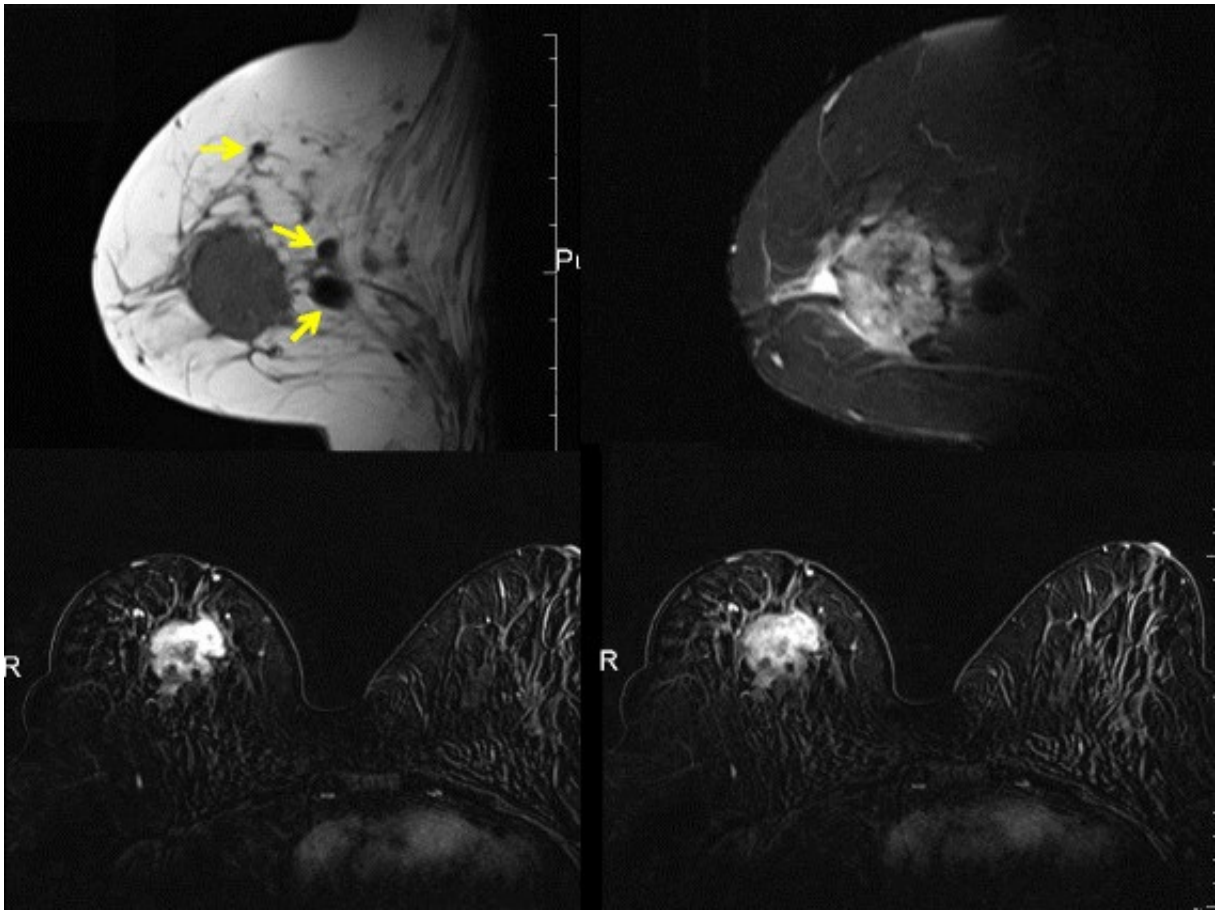

Figura 3.4: Imagens de uma ressonância magnética da mama
(Nakahori et al., 2015)

3.3 Sistemas de Apoio à Decisão Clínica

O médico tem a missão de interpretar um conjunto de dados, que muitas vezes não são estruturados, geralmente colhidos com o passar dos episódios clínicos. Os contínuos avanços na ciência da computação e suas tecnologias tecnologias têm sido cada vez mais utilizados na área da saúde. Com isso, são gerados uma enorme massa de dados nos hospitais, sempre que um paciente realiza algum tipo de exame médico, por exemplo, exame complementar, exame periódico, etc. Esses são relevantes para que profissionais de saúde possam fazer uma avaliação do estado clínico do paciente, muitas vezes, cruciais do processo de tomada de decisão do médico (Nogueira et al., 2012).

Para ajudar na tomada de decisão atualmente existem os Sistemas de Apoio à Decisão Clínica (SADC), que visam obter melhores desempenhos na resolução de problemas fornecendo recomendações específicas para cada paciente os quais são baseados em princípios de análises de decisões. Os Sistemas de Apoio à Decisão Clínica (SADC) são aplicações desenhadas para auxiliar os médicos na tomada de decisões de diagnóstico e de terapêutica nos cuidados a doentes (Silva et al., 2012). Um dos principais objetivos do SADC é reduzir a necessidade de atendimento especializado, muitas vezes demorado, a um determinado paciente. Isso é

possível através do fornecimento de ferramentas de apoio à decisão, bem como oferecer as evidências necessárias para que os profissionais de saúde possam classificar os achados clínicos. Reduzindo assim, o tempo de diagnóstico e prognóstico de pacientes (Silva et al., 2012).

No contexto de oncologia, também foram desenvolvidos diversos softwares que têm o objetivo de auxiliar no processo de tomada de decisão. Vale destacar que, há diversas razões para a construção deste tipo de sistema neste cenário, por exemplo, a incorporação de dados clinicopatológicos para predizer o tempo de sobrevida de um paciente após o diagnóstico de um algum tipo de tumor. Além disso, outra forma de utilização pode ser através do uso do perfil genético do paciente, para a partir dos dados laboratoriais e clínicos, obter informações tais como qual o método de tratamento individual, pois apenas um quarto dos pacientes com câncer responde de forma positiva aos medicamentos prescritos para os mesmos de forma generalizada (Banjar et al., 2017)

3.4 Aplicação de Inteligência Artificial no Prognóstico de Câncer de Mama

A inteligência artificial (IA) e a aprendizagem de máquina (AM) têm sido aplicados para classificação e diagnóstico do câncer de mama há quase duas décadas, porém não existem tantos estudos acerca da sua aplicação no contexto de prognóstico (Kourou et al., 2015).

Recentemente, foram aplicadas algumas técnicas de aprendizagem de máquina semi-supervisionada para o desenvolvimento de um modelo para progressão e capacidade de sobrevivência de pacientes diagnosticados com câncer de mama. A maioria, todavia, não utilizou dados importantes como status de receptores hormonais, por exemplo, HER2 ou Ki67. No entanto, outros modelos foram construídos através de modelos híbridos contendo dados de microarranjos ou imagens mamográficas (Ferroni et al., 2019).

Alguns dos sistemas de prognósticos desenvolvidos são o Adjuvant! Online e PREDICT. O sistema Adjuvante! Online foi desenvolvido utilizando os registros da Surveillance, Epidemiology and End-Results (SEER) para predizer os riscos de recorrência em 10 anos, mortalidade específica por câncer de mama ou por outras causas, levando em conta o benefício esperado das terapias sistêmicas pós operatórios (Ravdin et al., 2001). Já o PREDICT, foi criado usando informações dos registros de câncer de mama do Reino Unido, com o objetivo de prever o tempo de sobrevida entre 5 e 10 anos para pacientes diagnosticados com câncer de mama, de acordo com os benefícios esperados do tratamento de quimioterapia, terapia alvo e endócrino (Wishart et al., 2010).

3.5 Algoritmos para seleção de características

Podemos definir características, no contexto de aprendizagem de máquina (ou do inglês *machine learning* ou ML), podem ser definidas como àquelas propriedades de fenômeno observado. Geralmente são utilizadas em algum processo de identificação sistemática de padrões, através do uso de modelos de ML. Denominamos seleção de características como um processo que tem o objetivo de reduzir a quantidade de propriedades que são aplicadas em um modelo de ML. A etapa de redução de características pode ser feita através da análise e processamento das características, com a finalidade de reconhecer quais destas são as mais relevantes dentre todas as existentes (Sherer, 2018). Um exemplo de aplicação de seleção de características na área médica é na análise de genes de um microarranjo de amostras de célula de um determinado organismo de um paciente. Este tipo de dado contém milhares de variáveis, que apresentam uma certa correlação entre elas. Este tipo de dependência entre as propriedades, não fornece informação extra sobre a variável alvo, sendo assim geram ruídos para o modelo de aprendizagem de máquina utilizado. Sendo assim, eliminar este tipo de variável, pode resultar em uma melhoria no desempenho do resultado do modelo utilizado.

Portanto, faz-se necessário a definição de um critério para remoção de características que não tem um nível de relevância aceitável com a variável alvo. Por outro lado, vale a pena ressaltar que a remoção de características irrelevantes não tem relação com métodos de redução de dimensionalidade de um conjunto de dados, tais como o *Principal Component Analysis* (PCA). A eliminação de propriedades através de técnicas de seleção de características não cria novas propriedades. Pois uma vez que é selecionado um critério de seleção de características, é iniciado um procedimento para encontrar um subconjunto de características úteis dentro do conjunto analisado (Chandrashekar and Sahin, 2014).

O processo de seleção de características vem sendo pesquisado ao longo dos anos, tando do ponto de vista prático como metodológico, e sido aplicado em diversas áreas, tais como análise de dados bioinformáticas, reconhecimento de imagens, mineração de texto, seleção de genes que são catalisadores para diversas doenças, etc. Além disso, eles podem ser classificados de acordo com o padrão utilizado. Se forem analisados de acordo com os dados de treinamento, estes podem ser classificados como modelos supervisionados, não supervisionados e semi-supervisionados, porém estes termos não devem ser confundidos com a classificação de algoritmos de aprendizagem de máquina puramente. Por outro lado, se forem analisados de acordo os métodos de aprendizagem, podem ser classificados como filters, wrappers e embedded (Cai et al., 2018). A Figura 3.5 demonstra um diagrama sobre a categorização dos modelos de seleção de características.

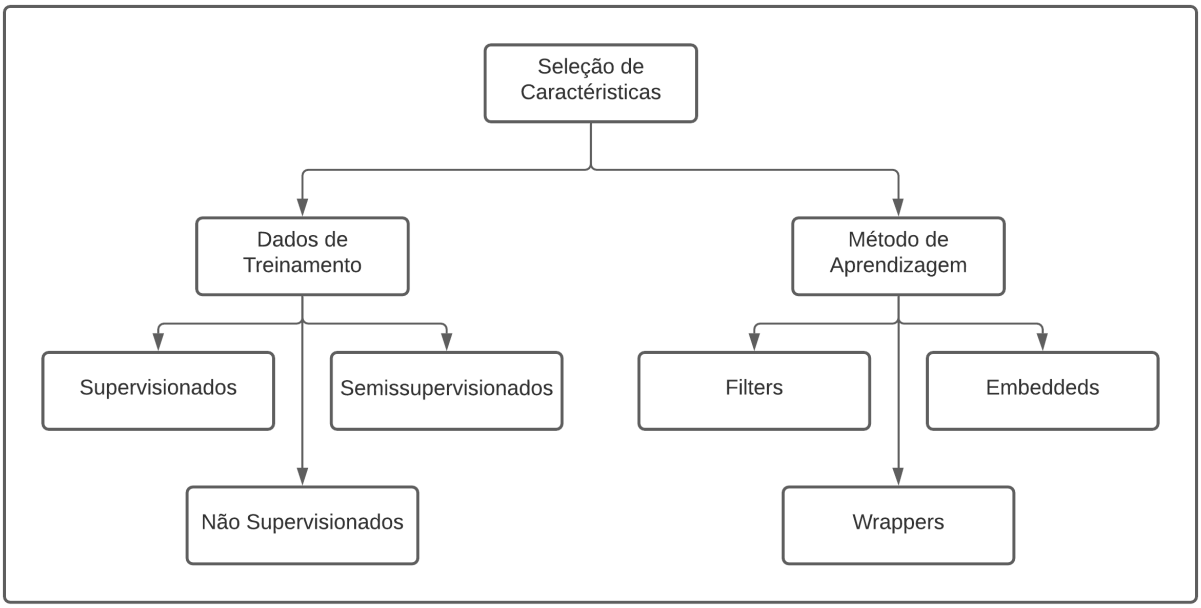

Figura 3.5: Categorias de seleção de características
(Sammut and Webb, 2017)

3.5.1 Padrão de classificação de acordo com os dados de treinamento

Modelos Supervisionados

Modelos deste tipo geralmente são utilizados em problemas de classificação. A disponibilidade de rótulos no conjunto de dados, permitem algoritmos supervisionados selecionar as características mais efetivas na distinção de instâncias de diferentes rótulos. Neste cenário, as características são selecionadas a partir conjunto de dados de treinamento. Dessa forma, não são utilizados todas as propriedades no treinamento do modelo de aprendizagem supervisionada, inicialmente seleciona-se um subconjunto de características, após isso aplica-se o conjunto de dados resultantes sobre o modelo de aprendizagem utilizado. Na etapa de seleção de características utilizou-se das informações das classes do conjunto de dados (Sammut and Webb, 2017). Alguns exemplos de algoritmos supervisionados são: Information Gain e Gain Ratio. Podemos verificar na Figura 3.6 o fluxo de seleção de características em modelos supervisionados.

Modelos Não Supervisionados

Já os modelos não supervisionados são usados em problemas de clusterização. Considera-se a estrutura do funcionamento deste tipo de modelo bastante parecido com o de modelos supervisionados, exceto o fato de que não há informações de classes nos dados utilizados. Sem este tipo de informação para definir a efetividade de uma característica, modelos não supervisionados têm que utilizar outro tipo de critério para analisar a relevância de uma característica. Neste

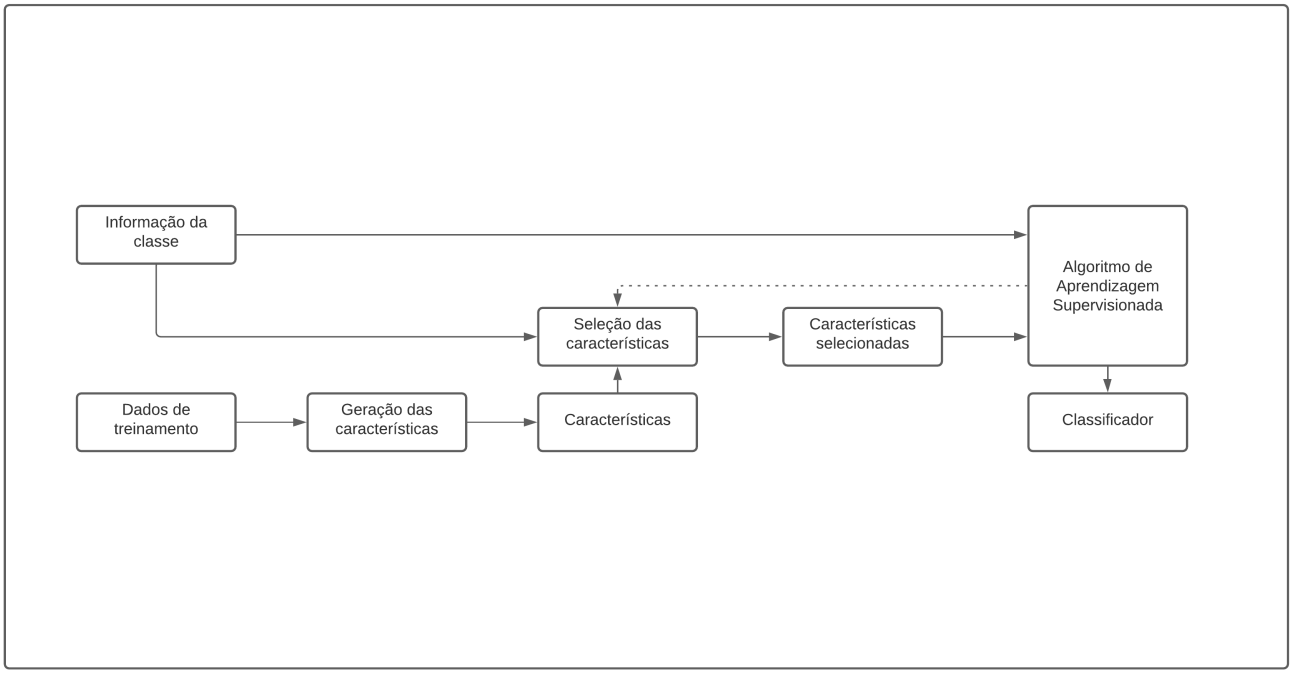

Figura 3.6: Seleção de características em modelos supervisionados

contexto geralmente utiliza-se um critério capaz de escolher uma característica que preserve a estrutura múltipla dos dados originais. Outro método utilizado para definir a relevância das propriedades é buscar indicadores de agrupamento por meio de algoritmos de agrupamento, após esta etapa, transformar a estrutura não supervisionada em uma supervisionada. Ainda segundo (Sammut and Webb, 2017) existem duas formas diferentes de utilizar algoritmos de agrupamento, a primeira é buscar indicadores de agrupamento e paralelamente selecionar as características supervisionados dentro de uma estrutura unificada. A segunda forma é, primeiro buscar o indicador de agrupamento, logo após, utiliza-se uma característica para adicionar ou remover outras características, por fim, repetir este processo até que um determinado critério seja satisfeito. Na Figura 3.7 podemos encontrar a estrutura de funcionamento de modelos não supervisionados.

Modelos Semi-supervisionados

Geralmente estes tipos de modelos são empregados em dados cujo apenas uma parte dos dados é rotulada. Neste cenário, a utilização de modelos supervisionados ou não-supervisionados podem não ser a melhor opção para seleção de características. Pois a seleção de características utilizando métodos supervisionados podem não ser capazes de selecionar as características mais relevantes, porque a quantidade de informações rotulados no conjunto de dados são insuficientes para representar a distribuição das características.

Já os métodos não supervisionados não usam informações rotuladas, o que é um ponto negativo, pois as poucas informações rotuladas existentes podem ter uma certa relevância na

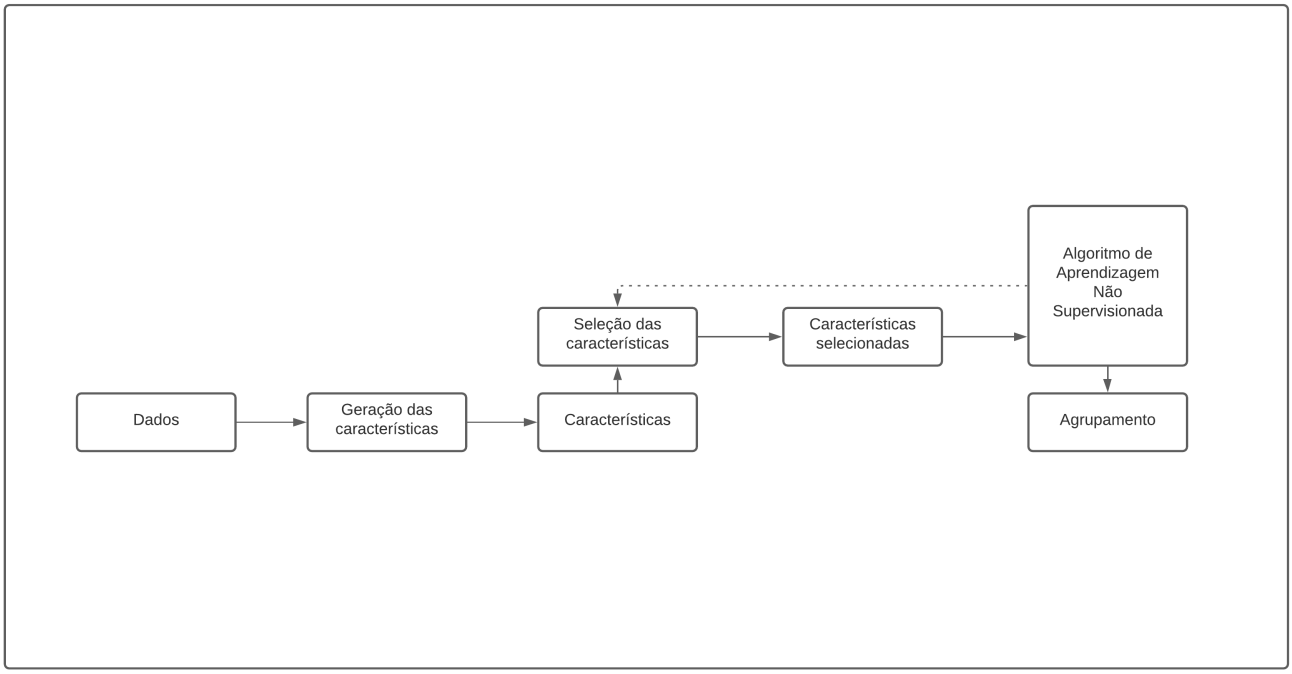

Figura 3.7: Seleção de características em modelos não supervisionados

determinação das características mais efetivas dentro do conjunto de dados. Por outro lado, modelos semi-supervisionados de seleção de características, podem fazer o melhor proveito dos dados rotulados e não rotulados, sendo a melhor escolha quando temos dados parcialmente rotulados. A maioria dos modelos semi-supervisionados existentes dependem da construção de uma matriz de similaridade a fim de selecionar as características que mais se assemelham a esta. Dessa forma, tanto as informações dos rótulos dos dados, quanto a medida de similidade dos dados são utilizados para construção da matriz de similaridade, para que dessa forma, as informações rotuladas possam fornecer aspectos discriminativos na seleção de características relevantes, enquanto os dados não rotulados são capazes de fornecer informações complementares (Sammut and Webb, 2017).

3.5.2 Filters

Métodos do tipo filter Figura 3.8 funcionam, inicialmente, estabelecendo um *ranking* das características utilizando algum critério preestabelecido. São executados em uma etapa anterior a aplicação do uso do modelo de ML, e são independentes do modelo aplicado no estudo. Este tipo de técnica pode ser classificado de acordo com os parâmetros de filtragem que empregam, tais como nível de dependência entre as propriedades ou grau de similidade (Urbanowicz et al., 2018a).

Porém, este tipo de modelo não leva em consideração o viés e as heurísticas dos algoritmos de aprendizagem de máquina. Dessa forma, podem ser que faltem recursos que são

relevantes para o algoritmo de aprendizagem alvo. Alguns dos critérios utilizados neste tipo de classificação das características são: a capacidade efetiva de separar amostras de diferentes classes considerando a variação de classes e a variância entre elas, além de analisar a dependência entre uma característica e o rótulo da instância analisada, verificar também a correção entre as classes dos dados e entre as características presentes. Porém, vale ressaltar que a principal desvantagem deste tipo de abordagem é que é ignorado totalmente o efeito do subconjunto de características selecionados na execução de algoritmos de classificação ou agrupamento. (Sammut and Webb, 2017). Exemplos deste tipo de algoritmo são o Information Gain, Chi-Square, Gain Ration e Relief-F .

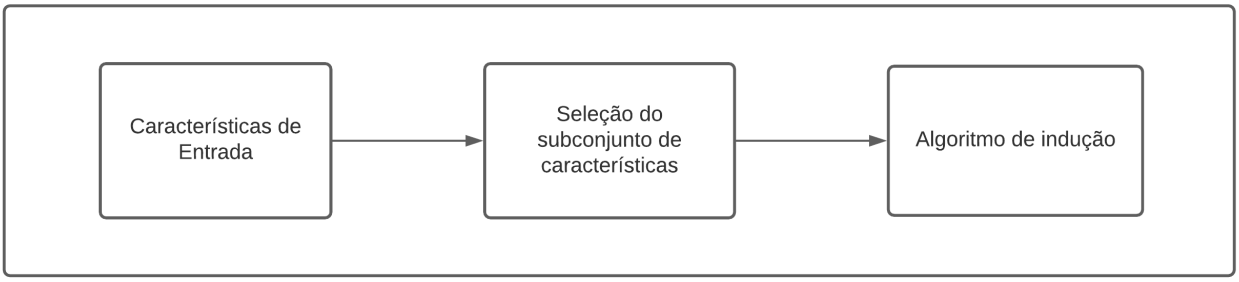

Figura 3.8: Modelo de seleção de características do tipo *filter*

3.5.3 Wrappers

Métodos wrappers Figura 3.9 funcionam criando uma quantidade de subconjuntos pré-definido de acordo com a quantidade total de características, em seguida calcula um score para cada subconjunto utilizando uma função objetiva específica. O subconjunto com a maior performance é considerado em detrimento dos demais subconjuntos. Geralmente este tipo de modelo demanda um maior esforço computacional em relação aos algoritmos filters (Singh, 2019).

O subconjunto de características ideal depende do viés e heurísticas específicas dos algoritmos de aprendizagem. Com base nisto, modelos wrappers usam algoritmos de aprendizagem específicos na avaliação da qualidade das características (Sammut and Webb, 2017).

Diferentemente de métodos do tipo filter, métodos wrappers estão relacionados ao modelo e aprendizagem de máquina aplicado, e é executado na fase de treinamento do modelo, caso seja utilizado um novo modelo de aprendizagem, deve ser refeito o processo de definição do subconjunto de características.

3.5.4 Embedded

Os métodos embedded podem ser definidos como a combinação dos métodos citados anterior. Com a vantagem de ser mais performático que os métodos wrappers, pois realiza a integração

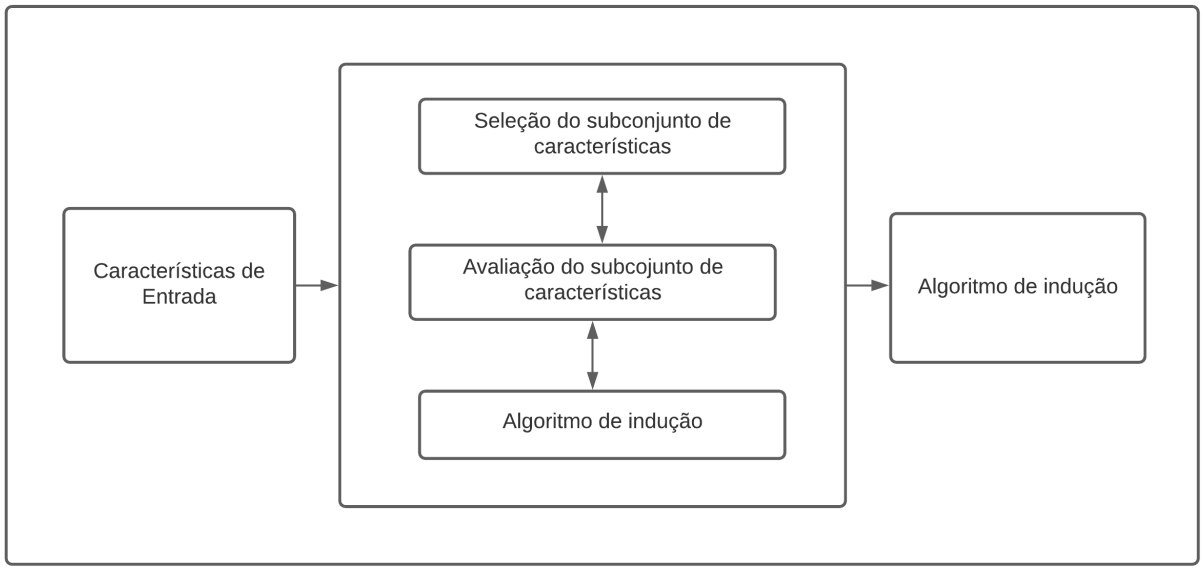

Figura 3.9: Modelo de seleção de características do tipo *wrapper*

de forma paralela entre o modelo e as características selecionadas. Isto pode ser realizada, por exemplo, através de uma função objetiva dividida em duas partes, primeiramente definindo um termo de adequação e, estabelecendo uma penalidade para um subconjunto com um número elevado de características (Singh, 2019).

3.6 Algoritmos de Aprendizagem de Máquina

Durante a sua evolução, a humanidade tem utilizado várias ferramentas para facilitar a forma como são feitas tarefas complexas. Como consequência, o cérebro humano foi responsável pela criação de diferentes tipos de máquinas. Estas por sua vez, tem facilitado bastante a vida humana, possibilitando a realização de várias tarefas de uma forma mais satisfatória, tais como viagens, tarefas industriais e também no campo da informática. Tratando-se de benefícios que a evolução trouxe para o contexto de computação, podemos citar a aprendizagem de máquina.

De acordo com (Samuel, 1959), Aprendizagem de Máquina é um subcampo da inteligência artificial que lida com algoritmos de computação que podem ser melhorados via dados de treinamento sem programação explícita, ou seja, ela é capaz de possibilitar os computadores a aprender sem ser explicitamente programado para tal. (Samuel, 1959) é reconhecido por ser um pioneiro no campo dos jogos de computador, inteligência artificial e aprendizado de máquina,

A Aprendizagem de Máquina é utilizada para que ensinar aos computadores como manipular dados de uma maneira mais eficiente e eficaz. Pois, para nós seres humanos muita vezes é difícil de visualizar e interpretar as informações contidas em um conjunto de massa de dados. Por isso, nestes casos a aplicação de de aprendizagem de máquina traz a possibilidade de extrair informações relevantes de um conjunto de dados, onde dificilmente nós humanos

conseguiríamos enxergar algo (Mahesh, 2020).

Algoritmos de Aprendizagem de Máquina podem ser classificados, de maneira geral, em três categorias: supervisionados; não-supervisionados; e aprendizagem por reforço.

3.6.1 Aprendizagem Supervisionada

A aprendizagem supervisionada pode ser definida como um tipo de aprendizagem ao qual uma função mapeia uma determinada entrada para uma saída correspondente de acordo com o conjunto de dados em análise. Algoritmos de aprendizagem supervisionada são aqueles que necessitam de alguma assistência externa. O conjunto de dados de entrada geralmente é dividido um subconjunto para treinamento do modelo e um outro conjunto para testes, onde através deste processo é feito a validação da função definida para o mapeamento entre entrada-saída. A ideia principal é que este tipo de algoritmo aprenda algum tipo de padrão a partir de um conjunto de dados prévio, através deste padrão conseguir inferir alguma informação relevante para o propósito de aplicação do modelo de aprendizagem de máquina (Mahesh, 2020). Alguns dos algoritmos mais famosos neste contexto são: Árvore de Decisão, Naive Bayes, kNN, Regressão Linear, Regressão Logística e Floresta Aleatória.

3.6.2 Aprendizagem Não-Supervisionada

Diferentemente da classe de algoritmos supervisionados, este tipo algoritmo não há um mapeamento explícito entre entrada e saída dos dados. Algoritmos desta natureza, tem a capacidade própria de aprender padrões e estruturas de um conjunto de dados. Neste caso, uma vez que o algoritmo aprende algum determinado padrão sobre os dados, quando são introduzidos novos dados, o mesmo utiliza o conhecimento prévio para reconhecer classes entre dos dados. Podemos citar como algoritmos mais famosos neste cenário como sendo: Análise de Componente Principal, K-Means e Clusterização Hierárquica.

3.6.3 Aprendizagem por Reforço

A Aprendizagem por reforço é uma área da aprendizagem de máquina que tem o objetivo de analisar como os agentes de software devem tomar as decisões em um ambiente, com a finalidade de maximizar alguma noção de recompensa cumulativa (Mahesh, 2020).

4

Algoritmo Proposto

Neste capítulo apresentaremos o algoritmo proposto para solucionar o problema da identificação de características biológicas mais relevantes no prognóstico da doença de câncer de mama. Primeiramente, apresentaremos uma visão geral do algoritmo Relief-F. Em seguida, abordaremos o funcionamento do algoritmo Gain Ratio. Por fim, apresentaremos a abordagem utilizada para criação do algoritmo híbrido proposto neste trabalho.

4.1 Relief-F

Esta seção apresenta uma simples visão do funcionamento do algoritmo Relief-F.

O algoritmo Relief-F pertence à família de algoritmos Relief. Os algoritmos desta família podem ser divididos em três grupos: o algoritmo básico Relief; o algoritmo Relief-F e o algoritmo RRelief-F. O algoritmo básico é limitado a ser utilizado na classificação de problemas com duas classes. Já o algoritmo Relief-F é considerado uma extensão do algoritmo básico, e consegue se sair bem no cenário em que temos multiclasses. Por fim, o algoritmo RRelief-F é uma adaptação do algoritmo Relief-F para problemas de regressão (Robnik-Šikonja and Kononenko, 2003a).

Uma das vantagens do algoritmo em relação aos outros algoritmos de seleção de características é que, enquanto outros modelos assumem uma independência condicional entre os atributos do conjunto de dados para estimar a qualidade. Já os algoritmos da família Relief-F não fazem esta suposição, e são bastante eficientes na seleção das propriedades, levando em consideração na avaliação da qualidade das propriedades a dependência existente entre estas (Robnik-Šikonja and Kononenko, 2003a).

A ideia original do algoritmo Relief, demonstrada no algoritmo 1, é estimar a qualidade dos atributos de um conjunto de dados de acordo com a proximidade entre as instâncias pertencentes ao mesmo. Para realizar esta distinção, primeiramente é selecionada uma instância

aleatória R_i (linha 3), em seguida são procurados os vizinhos mais próximos pertencentes a mesma classe, chamando o vizinho que tem maior proximidade de H, e outro vizinho de classe diferente, é chamado de M (linha 4). Após isto, atualiza-se a estimativa da qualidade $W[A]$ para todos os atributos A dependendo de seus valores R_i, M e H (linhas 5 e 6). Observando que, se as instâncias R_i H tiverem valores diferentes do atributo A, então este atributo separa duas instâncias pertencentes a mesma classe, então diminui-se a estimava da qualidade de $W[A]$. Por outro lado, se A separa duas instâncias com valores de classes diferentes, então aumenta-se a estimativa da qualidade de $W[A]$, o que é desejável. Dessa forma, todo este processo é repetido m vezes, onde m é um parâmetro definido pelo usuário (Robnik-Šikonja and Kononenko, 2003a).

O presente trabalho apresenta como entregável, o algoritmo em questão

Algorithm 1: Relief Original

Input : Para cada instância de treinamento, um vetor de valores de atributos e a classe valor

Output: O vetor W de estimativas das qualidades dos atributos

1 set all weights $W[A] := 0.0$;
2 **for** $i := 1$ ***to*** m **do**
3 randomly select an instance R_i;
4 find nearest hit H and nearest miss M;
5 **for** $A := 1$ ***to*** a **do**
6 $W[A] := W[A] - diff(A, R_i, H)/m + diff(A, R_i, M)/m$;
7 **end**
8 **end**

A função $diff(A, I_1, I_2)$ calcula a diferença existente das duas instâncias I_1 e I_2 para o atributo A. Para atributos nominais ela é definida como:

$$diff(A, I_1, I_2) = \begin{cases} 0, & value(A, I_1) = value(A, I_2). \\ 1, & \text{otherwise.} \end{cases}, \tag{4.1}$$

e para atributos numéricos:

$$diff(A, I_1, I_2) = \frac{|value(A, I_1) - value(A, I_2)|}{max(A) - min(A)} \tag{4.2}$$

O algoritmo original Relief consegue trabalhar bem em conjunto de dados com atributos nominais e numéricos. Porém, ele é limitado a funcionar em conjuntos com duas classes. Para casos de dados com multi-classes, é utilizada sua extensão Relief-F. Cujo algoritmo é descrito

a seguir:

Algorithm 2: Relief-F

Input : Para cada instância de treinamento, um vetor de valores de atributos e o valor da classe

Output: O vetor W de estimativas das qualidades dos atributos

1 set all weights $W[A] := 0.0$;
2 **for** $i := 1$ ***to*** m **do**
3 randomly select an instance R_i;
4 find k nearest hit H_j;
5 **for** *each class* $C \neq$ *class* (R_i) **do**
6 from class C find k nearest misses $M_j(C)$;
7 **end**
8 **for** $A := 1$ ***to*** a **do**
9 $W[A] := W[A] - \sum_{j=1}^{k}$ *diff*(A, R_i, H_j)$/(m.k) + \sum_{C \neq (R_i)} (\frac{P(C)}{1-P(class(R_i))} \sum_{j=1}^{k}$ *diff*(A, $R_i, M_j(C)$)$/(m.k)$);
10 **end**
11 **end**

O algoritmo Relief-F não está limitado a problemas com classes binárias, ele é mais robusto que o original e pode trabalhar bem em dados com ruídos. Seu funcionamento é similar ao algoritmo original, inicialmente é selecionada uma instância R_i aleatória (linha 3), depois ocorre a busca pelo k vizinho mais próximo pertencente a mesma classe, chamado de H_j (linha 4), também é buscado o vizinho mais próximo pertencente a outra classe, este chamado de $MJ_j(C)$ (linhas 5 e 6). A atualização da estimativa $W[A]$ da qualidade de todos atributos A depende dos valores para R_i, H_j e $M_j(C)$ (linhas 7, 8 e 9). Para trabalhar com dados faltantes é necessário mudar a função *diff*. Valores faltantes são tratados utilizando probabilidade. Calcula-se a probabilidade de duas instâncias que apresentam valores diferentes para um determinado atributo condicionado ao valor da classe da seguinte maneira:

se uma instância tem algum valor desconhecido:

$$\textit{diff}(A, I_1, I_2) = 1 - P(value(A, I_2)|class(I_1)) \tag{4.3}$$

se ambas instância têm algum valor desconhecido:

$$\textit{diff}(A, I_1, I_2) = 1 - \sum_{V}^{\#values(A)} (P(V|class(I_1)) \times P(V|class(I_1))) \tag{4.4}$$

Utilizam-se frequências relativas do conjunto de treinamento para realizar a aproximação de probabilidades condicionais (Robnik-Šikonja and Kononenko, 2003a).

4.2 Gain Ratio

Nesta seção apresentaremos uma breve introdução do funcionamento da medida seleção de característica *gain ratio*. Um dos modelos de aprendizagem de máquina mais utilizados é o de árvore de decisão. Ele é caracterizado como uma estrutura simples, onde nós não-terminais representam testes sobre um conjunto de atributos, e os nós terminais representam as saídas do problema. Neste contexto, o conceito de ganho de informação é definido como uma medida utilizada para selecionar a ordem dos atributos testes da árvore de decisão. Esta medida dá preferência à utilização de atributos que têm um grande número de valores. Existem vários tipos de algoritmos de árvore de decisão, o mais popular é o ID3. Porém, ao passar dos anos surgiram novas versões de algoritmo de indução, como por exemplo , o algoritmo C4.5. A figura 4.1 ilustra uma árvore de decisão gerada a partir do funcionamento do algoritmo Gain Ratio. Este algoritmo, que é tido como o sucessor do ID3, utiliza uma extensão de ganho de informação, denominada de *gain ratio*. Que tem como objetivo diminuir o viés da seleção de atributos com um alto número de valores. As informações necessárias para classificar uma determinada amostra, pode ser definida da seguinte forma:

$$I(S) = -\sum_{i=1}^{m} p_i log_2(p_i) \tag{4.5}$$

Onde S é definido é constituído de s amostras de dados com m classes distintas, e p_i é a probabilidade de uma amostra arbitrária pertencer a classe C_i e calcula através de s_i/s. Considerando que um atributo A tem valores distintos v. E s_{ij} seja o número de amostras das classes c_i em um subconjunto s_j. S_j contém as amostras em s que tem o valor a_j de A. Dessa forma, a entropia, ou informação esperada baseada na partição do conjunto A em subconjuntos é dada por:

$$I(S) = -\sum_{i=1}^{m} I(S)\frac{s_{1i} + s_{2i} + ... + s_{mi}}{s} \tag{4.6}$$

O cálculo do ganho de informação a partir da partição do conjunto A é definido como:

$$Gain(A) = I(S) - E(A) \tag{4.7}$$

O algoritmo C4.5, que utiliza o *gain ratio*, aplica uma normalização do ganho de informação utilizando uma informação definida como:

$$SplitInfo_A(S) = -\sum_{i=1}^{v}(|S_i|/|S|)log_2(|S_i|/|S|) \tag{4.8}$$

Este valor representa a informação produzida através da divisão do conjunto de dados S em v partições correspondentes a n resultados de testes realizados sobre o atributo A. O *gain ratio* é definido como:

$$GainRatio(A) = Gain(A)/SplitInfo_A(S) \tag{4.9}$$

Sendo assim, o atributo com a maior taxa de ganho é selecionado como atributo teste da árvore de decisão (Karegowda et al., 2010a).

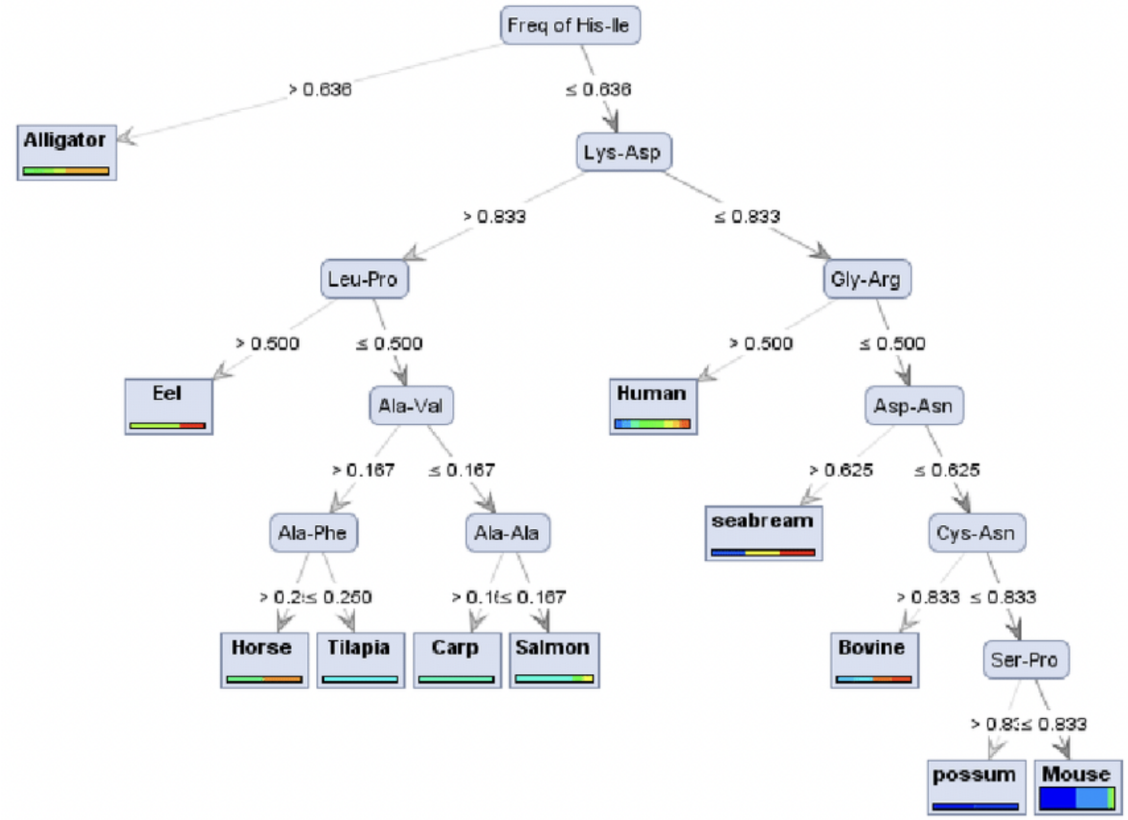

Figura 4.1: Árvore de decisão induzida por algoritmo de árvore de decisão executada em variáveis numéricas utilizando critério do algoritmo Gain Ratio
(Ebrahimi, 2011)

4.3 Algoritmo Híbrido

Um algoritmo híbrido pode ser definido como a combinação de dois ou mais algoritmos afim de resolver o mesmo problema. Segundo (Silva et al., 2020) alguns dos algoritmos mais utilizados para o prognóstico da doença de câncer são os algoritmos Relief-F e Gain Ratio. Levando este fator em consideração, a ideia proposta é realizar a combinação destes algoritmos e criar um novo modelo híbrido. O novo algoritmo consiste de realizar a seleção de característica sobre um conjunto de dados selecionando inicialmente um conjunto de atributos utilizando o algoritmo Gain Ratio, a partir do resultado desta primeira seleção, aplicar o algoritmo Relief-F, afim de

refinar a seleção de atributos.

Algorithm 3: Algoritmo Hibrido

Input : Um vetor de valores de atributos e o valor da classe

Output: Um subconjunto de dados contendo os atributos com as melhores estimativas de qualidade

```
for each feature f in dataset D do:
calculate the gain ratio (GR) of feature f;
select n best features to obtain subset M1;
set all weights W[A] := 0.0 for each feature A from subset M1;
for i := 1 to m do
    randomly select an instance R_i;
    find k nearest hit H_j;
    for each class C ≠ class (R_i) do
        from class C find k nearest misses M_j(C);
    end
    for A := 1 to a do
        W[A] := W[A] − Σ_{j=1}^{k} diff(A, R_i, H_j)/(m.k) + Σ_{C≠(R_i)} ( P(C)/(1−P(class(R_i))) ) Σ_{j=1}^{k} diff(A,
          R_i, M_j(C))/(m.k));
    end
end
select p best features from vector W to obtain subset M2;
output M2;
```

A figura 4.2 contém a representação do fluxo de execução do algoritmo proposto.

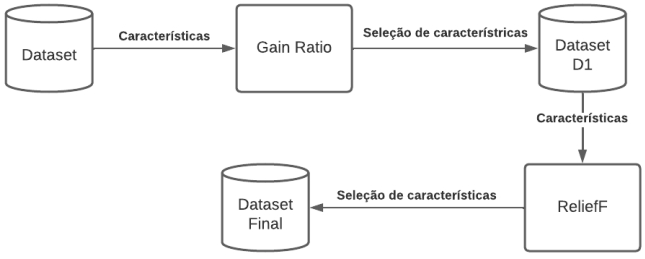

Figura 4.2: Fluxo de execução do algoritmo proposto (Ebrahimi, 2011)

A implementação do algoritmo proposto foi realiza através a utilização do framework Orange. Orange é um ferramente bastante utilizada para realização de análise de dados por meio de scripts na linguagem de programação Python e também na forma de programação visual (Demšar et al., 2013).

Posto isto, vale ressaltar que a implementação do algoritmo em questão foi realizada através da interface gráfica do framework Orange. Através deste tipo de programação é possível selecionar os componentes necessários para implementação do modelo que se deseja cons-

truir, o programa dispõe de vários modelos de aprendizagem de máquina, também dispõe de aspectos estatísticos sobre o conjunto de dados, bem como modelos de seleção de características, dentre os quais estão presentes os modelos de seleção utilizados no trabalho proposto. A utilização das ferramentas disponíveis no Orange, dar-se através da construção de um fluxo desejável, para construir o fluxo é preciso apenas arrasta os componentes que se deseja trabalha para a área de trabalho da ferramenta Orange. Sendo assim, foi assim que foi implementado o modelo híbrido proposto neste trabalho, uma vez definido o fluxo do algoritmo, foram utilizados os componentes pertinentes para construção do mesmo. A figura 4.3 traz o fluxo utilizado para implementar o algoritmo de seleção de características objeto deste trabalho. A região destacada em vermelho representa o fluxo de predição utilizando apenas o modelo Gain Ratio como seleção de características sobre os modelos de aprendizagem de máquina Redes Neurais, Floresta Aleatória, kNN e Naive Bayes. Já a região em amarelo representa a predição de mortalidade de câncer de mama aplicando apenas o modelo Relief-F como seletor de características. Por fim, a região em verde demonstra o fluxo de predição de mortalidade aplicando o modelo proposto como seletor de características.

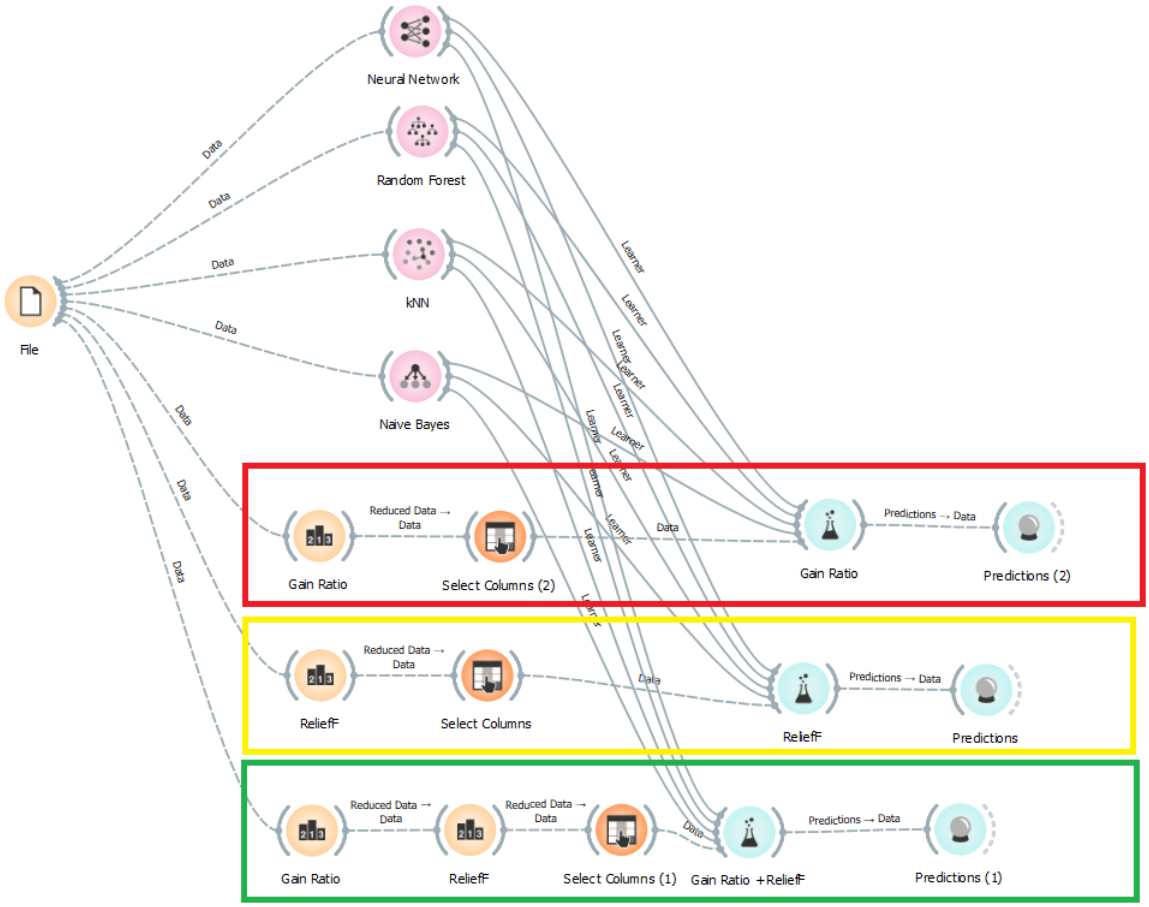

Figura 4.3: Implementação do algoritmo proposto utilizando o framework Orage (Demšar et al., 2013)

A máquina utilizada para realização do experimento apresenta a seguinte configuração:

processador Intel 8th generation (modelo 8400); 32Gb de memória RAM (2400Mhz); e sistema operacional Windows 11.

4.3.1 Algoritmo em termos de entregável e utilização

Podemos definir como entregável da trabalho em questão, um modelo de aprendizagem de máquina que pode ser utilizado por cientistas de dados na construção de sistemas de apoio a decisão médica. A sua implementação pode ser realizada através de bibliotecas bastante conceituadas e utilizadas no meio acadêmico e no mercado profissional, a saber: Orange (Demšar et al., 2013) e Sklearn (Pedregosa et al., 2011).

4.4 Algoritmos de Aprendizagem de Máquina utilizados para validação das características selecionadas

4.4.1 K-Nearest Neighbour (k-NN)

K-Nearest Neighbour é um dos métodos de classificação mais simples. O algoritmo representa cada exemplo de treinamento como um ponto num espaço, chamado de espaço de características. Os elementos são classificados com base em sua proximidade em relação aos demais exemplos nesse espaço. Para utilizar o k-NN é necessário definir uma métrica para calcular a distância entre os elementos.

Quando uma nova instância é introduzida, k-NN encontra os k vizinhos mais próximos utilizando a medida de distância adotada. O rótulo da nova instância é determinado pela classe mais comum entre os seus k vizinhos mais próximos (k é um inteiro positivo, normalmente pequeno). Se k = 1, então o objeto é simplesmente atribuído à classe do vizinho mais próximo Hand et al. (2002).

Neste estudo, a métrica para calcular a distância entre os exemplos de treinamento é definida pela distância Euclidiana. O valor para k, número de vizinhos, foi determinado inicialmente igual 1. Depois, esse valor foi aumentado para 2, quando obteve um resultado ligeiramente melhor. Outros valores de k também foram testados, mas não houve variação considerável na taxa de erro do classificador.

4.4.2 Naïve Bayes

Naïve Bayes é um classificador probabilístico que utiliza o Teorema de Bayes através da suposição de que cada atributo é independente de todos os outros atributos. A construção do classificador é bem simples, por isso é chamado de naive (ingênuo). A probabilidade de um elemento ser rotulado com uma classe pode ser obtida pelo produto das probabilidades condicionais individuais de cada atributo dada a classe.

Apesar de sua concepção ingênua e de sua simplicidade, classificadores Naïve Bayes têm sido aplicados de forma satisfatória em muitas situações complexas do mundo real Domingos and Pazzani (1993).

4.4.3 Floresta Aleatória

O método de Aprendizagem de Máquina Floresta Aleatória é bastante utilizado para classificação e regressão. Além disso, é um algoritmo que pertence a classe dos algoritmos supervisionados de Aprendizagem de Máquina. Ele é um dos algoritmos mais utilizados pela sua simplicidade e diversidade, e também pelos bons resultados apresentados mesmo sem a utilização de hiper-parâmetros. Além disso, ele pode ser utilizado tanto para tarefas de classificação ou de regressão (Breiman, 2001).

A ideia principal deste algoritmo é construir uma coleção de árvores de decisão com uma variação controlada. Utilizando ensacamento, cada árvore de decisão no conjunto é construída utilizando uma amostra com substituição a partir dos dados de formação. Cada árvore do conjunto atua como classificador de base para determinar a etiqueta de classe de uma instância não etiquetada. Isto é feito através de votação por maioria, em que cada classificador dá um voto para a sua etiqueta de classe prevista, depois a etiqueta de classe com mais votos é utilizada para classificar a instância (Fawagreh et al., 2014).

4.4.4 Redes Neurais Artificiais

As redes neurais artificiais podem ser definidas como um conceito em computação que visa trabalhar com o processamento de dados de uma forma semelhante ao cérebro humano. O cérebro é pensado como um processador altamente complexo que executa o processamento em paralelo. Para o fazer tal tarefa, organiza a sua estrutura, mais conhecidos como neurônios, de tal forma que estes executam o processamento necessário. Isto é feito a uma velocidade extremamente alta, e não há computador no mundo capaz de fazer o que o cérebro humano faz.

Em redes neurais artificiais, a ideia é executar o processamento de informação usando como princípio a organização dos neurônios no cérebro. Uma vez que o cérebro humano é capaz de aprender e tomar decisões baseadas na aprendizagem, as redes neuronais artificiais devem fazer o mesmo. Assim, uma rede neural pode ser interpretada como um esquema de processamento capaz de armazenar conhecimento baseado na aprendizagem (experiência) e disponibilizar este conhecimento para a aplicação em questão (Mahesh, 2020).

5

Resultados e Discussões

Esta seção apresenta o conjunto de dados utilizados no experimento, bem como traz uma avaliação da eficácia do algoritmo proposto. Também apresentaremos uma comparação das métricas de avaliação dos modelos de aprendizagem de máquina sem aplicar o algoritmo proposto e após a aplicação do mesmo, as métricas utilizadas para avaliar os modelos foram acurácia, f1, precisão e sensibilidade. Para realizar a comparação foram utilizados os seguintes modelos de aprendizagem de máquina: (i) Redes Neurais, (ii) Random Forest (iii) Naive Bayes (iv) KNN

5.1 Conjunto de Dados

No experimento realizado neste trabalho, foi utilizado o *dataset* denominado The Molecular Taxonomy of Breast Cancer International Consortium (Curtis et al., 2012). Este conjunto de dados faz parte de um projeto desenvolvido por universidades do Reino Unido e Canadá, ao qual contém perfis de microarrays de mDNA de aproximadamente 2000 amostras obtidas através da plataforma Illumina HT-12 v3. Inicialmente, o total de amostras foi dividido em dois subconjuntos: (i) treinamento (997 amostras); (ii) e testes (989 amostras). Os dados foram coletados pelo Professor Samuel Aparicio do Centro de Câncer da Universidade de Columbia no polo do Canadá e o Professor Carlos Caldas do Instituto de Pesquisa de Cambridge, publicado na Nature em 2016 (Curtis et al., 2012). As amostras contêm dados clínicos e genéticos dos pacientes, e foi baixada do website cBioPortal. Neste contexto, as amostras foram subdivididas em cinco categorias: a) luminal A, luminal B, enriquecimento HER2, tipo normal e tipo basal, de acordo com a metodologia PAM50 (Parker et al., 2009). Vale ressaltar que o estudo em questão foi aprovado pelo Comitê de Revisão Institucional (Curtis et al., 2012).

5.2 Resultados

Após a obtenção dos dados experimentais, a primeira etapa necessária para executar o experimento foi a normalização dos dados. Após a finalização desta etapa, foi implementado o algoritmo híbrido proposto na ferramenta Orange (Demšar et al., 2013). Esta etapa consistiu em estabelecer a característica alvo do conjunto de dados, neste caso, definiu-se a característica que diz respeito se o paciente viveu ou morreu após ser diagnosticado com a neoplasia câncer de mama.

Após a seleção das características relevantes para o contexto analisado, foi iniciado o processo de treinamento dos modelos utilizados como base no experimento realizado. Os modelos usados foram: Random Forest, Neural Network, Naive Bayes e kNN.

Como a ferramenta utilizada para implementação dos modelos não dispõe de uma ferramenta de boosting dos hiper-parâmetros, foi necessário realizar de forma manual o experimento dos valores mais apropriados para obter os melhores resultados na etapa de treinamento de cada modelo.

Além disso, no processo de treinamento de cada modelo de aprendizagem de máquina utilizado no experimento, foi utilizado o método de validação cruzada k-fold para executar a validação cruzada. Na validação cruzada k-fold, você divide os dados de entrada em subconjuntos de dados k (também chamados de folds). Você treina um modelo de ML em todos, menos em um (k-1) dos conjuntos de dados e, em seguida, avalia o modelo no conjunto de dados que não foi usado para treinamento. Esse processo é repetido k vezes, com um subconjunto diferente reservado para avaliação (e excluído do treinamento) a cada vez (Refaeilzadeh et al., 2009). No experimento em questão foi utilizado o valor para k = 10.

No modelo Random Forest, precisamos qual o melhor valor do parâmetro k. Desta forma, foram testados vários valores para este parâmetro. Como pode ser observado na figura a seguir, a acurácia do modelo tem um valor aproximadamente constante para um valor de k maior ou igual a 15. Portanto, como não houve um ganho considerável na acurácia do modelo, e com valores muito grande para k, impactaria na diminuição da performance do modelo, escolheu-se o valor de k como sendo 15.

Já no modelo Neural Network os parâmetros que trouxeram os melhores resultados de avaliação do modelo foram: número de camadas ocultas igual a 100, Relu como parâmetro de ativação, o solucionador utilizado foi Adam, o taxa de regularização alfa igual a 0,0001 e o número máximo de iterações como sendo 200. A configuração de parâmetros utilizados no modelo kNN, definiu-se como métrica Manhattan e k igual a 5.

Para determinar o número de características que trazia os melhores resultados de validação para todos os modelos de aprendizagem utilizados no experimento, foi necessário testar vários valores para o número total de características selecionadas, ou seja, realizamos a primeira etapa do algoritmo proposto através da seleção de características pelo modelo Gain Ratio, em seguida foi feita uma nova seleção de características utilizando o algoritmo RelieF.

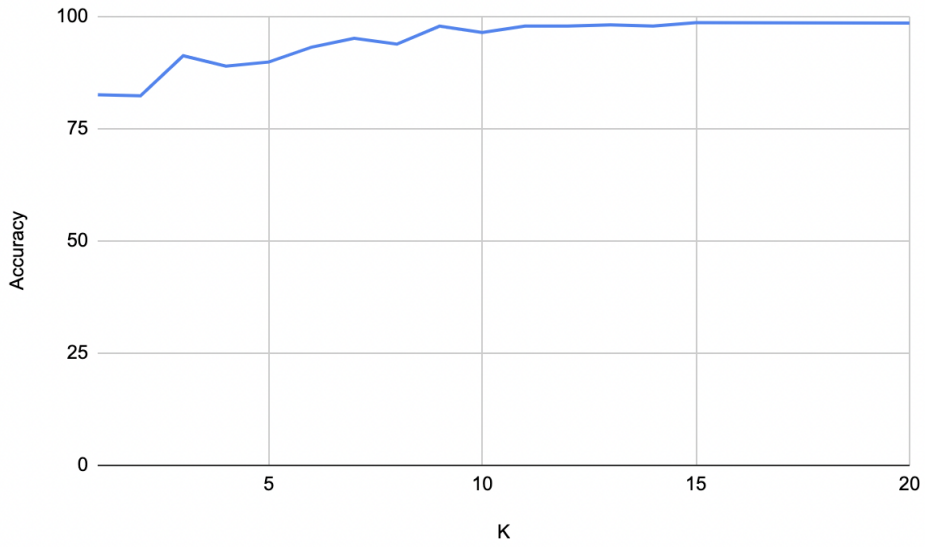

Figura 5.1: Variação da acurácia para diferentes valores de k sobre o modelo proposto

Foram feitos três testes iniciais, de um total de 692 características contidas no *dataset*, selecionamos as primeiras 500 melhores avaliadas pelo modelo Gain Ratio, depois selecionamos as 300 melhores ranqueadas pelo modelo Relief-F. O segundo teste filtrou as 300 melhores *features* através do Gain Ratio, em seguida as 200 melhores usando o Relief-F. Por fim, selecionamos as 300 melhores características segundo a avaliação do Gain Ratio e, logo após, selecionamos as 100 melhores características determinadas pelo Relief-F. Os melhores resultados obtidos foram obtidos através da utilização da quantidade de características determinadas pelo segundo teste, ou seja, selecionando as 200 melhores características do total contido no *dataset* original. As 25 características melhores qualificadas segundo cada algoritmo Gain Ratio podem ser verificadas na figura 5.2, já na figura 5.3 podemos visualizar as 25 características com os melhores valores qualitativos para o algoritmo Relief-F, e a figura 5.4 demonstra também a quantidade de 25 melhores características para o modelo proposto.

As figuras a seguir trazem uma comparação entre os resultados obtidos para cada abordagem de seleção de características, bem como o resultado de avaliação de cada modelo usando o conjunto de dados original. Sendo assim, a figura 5.5 demonstra os resultados para a métrica acurácia, podemos observar que o gráfico em questão trás os resultados de predição para as características selecionadas por três modelos diferentes de seleção de características, o primeiro é o Gain Ratio, o segundo o Relief-F e o terceiro é o modelo proposto, ou seja, a composição entre os dois algoritmos anteriores. Além disso, também traz a o resultado da métrica acurácia para os quatro modelos de aprendizagem utilizados no experimento, são eles: Redes Neurais, Floresta Aleatória, kNN e Naive Bayes. Através do gráfico podemos verificar que o modelo proposto trás melhores resultados para todos os modelos preditivos, com exceção do modelo Naive Bayes, que por sua vez apresenta um leve melhor resultado que os outros quando aplicado o modelo de seleção de características Relief-F. Mas isto é totalmente plausível, pois

tanto Relief-F como o Naive Bayes são modelos que fazem suposições probabilísticas, logo isso colabora para que o modelo Naive Bayes tenha melhores resultados quando o modelo de seleção de características utilizado é o Relief-F. Já a figura 5.6 traz os resultados para a métrica precisão, por outro lado na figura 5.7 é possível visualizar os resultados para a métrica sensibilidade, e por fim os resultados para a métrica F1 podem ser vistos na figura 5.8, os três últimos gráficos trazem resultados semelhantes aos obtidos para a métrica de avaliação acurácia.

5.3 Discussões

O modelo de seleção Gain Ratio é a modificação do algoritmo de Information Gain, que tem o objetivo de reduzir o viés causado por este último. Porém, uma de suas grandes desvantagens é que ele calcula um peso para um recurso sem examinar outros recursos disponíveis. Se os recursos forem dependentes, isso geralmente não será refletido em seus pesos. Um recurso que contém algumas informações sobre a classe de classificação por conta própria, mas nenhum quando outro recurso mais informativo estiver presente, receberá um peso diferente de zero. Os recursos que contêm pouca informação sobre a classe de classificação receberão um peso pequeno, mas um grande número deles ainda pode anular recursos mais importantes. Esses dois problemas terão uma influência negativa na precisão da classificação, principalmente quando houver muitos recursos disponíveis (Karegowda et al., 2010b).

Relief-F é uma extensão do algoritmo Relief. Como mencionado anteriormente, Relief é um algoritmo que define um peso para cada *feature*, este cálculo é feito levando em consideração a correlação com outras *features* e suas respectivas classes alvo. Através de um threshold, *features* com valor abaixo do definido são excluídos do conjunto de dados. Relief-F foi desenvolvido para tratar de conjuntos de dados com classes não binárias (Robnik-Šikonja and Kononenko, 2003b).

Porém, quando o algoritmo Relief-F é utilizado puramente como método de seleção de características, traz consigo o problema de não fazer uma boa avaliação qualitativa de características que têm uma certa dificuldade de distinção, ou seja, causando uma ineficiência de avaliação de classes minoritárias (Urbanowicz et al., 2018b).

Sendo assim, o modelo proposto tem o objetivo de diminuir este problema através da utilização do algoritmo Gain Ratio, pois este consegue fazer uma boa avaliação para classes minoritárias. Resultando na obtenção de um conjunto de características que representam de fato aquelas que têm maior relevância para a predição a ser realizada, e que trazem uma maior interpretatividade para os dados que serão analisados por um especialista da saúde que irá analisar os dados trabalhados pelo modelo de IA. Além disso, o algoritmo apresentado tenta diminuir o impacto dos pontos negativos de cada modelo em separado, criando um novo modelo que é capaz de levar em consideração as características que têm dependência entre si, bem como apresentar boas avaliações qualitativas das características mesmo em *datasets* como um nú-

mero muito elevado de características, que o caso de *datasets* contendo informações genéticas dos pacientes.

Portanto, vale destacar que o modelo demonstrou melhores resultados nas avaliações das métricas de sensibilidade e precisão. Ademais, esses resultados podem ter uma enorme relevância quando se trata de uma predição sobre o impacto que o câncer de mama pode ter sobre um paciente, pois estas métricas destacam o quanto o modelo pode diminuir a quantidade de falsos positivos bem como a diminuição de falsos negativos. Sendo assim, o profissional de saúde pode ter uma atenção mais adequada para cada paciente de acordo com as possibilidades de quadro clínico do paciente de acordo com os valores obtidos através de exames clínicos e genéticos realizados pelo mesmo. Porém, vale ressaltar que para validar a importância dos resultados apresentados, é necessária uma avaliação das características selecionadas por um profissional de saúde da área, pois este tem a capacidade de confirmar se os dados apresentados são coerentes ou não.

Por outro lado, os resultados também mostraram que as características selecionadas pelo modelo Relief-F demonstram serem mais satisfatório em relação aos demais modelos para o modelo de aprendizagem de máquina Naive Bayes, isso acontece pelo fato deste algoritmo trabalhar utilizando um modelo probabilístico e faz uso de suposição de independência entre as variáveis preditoras, assim também como a abordagem da seleção de características do Relief-F, resultando em um melhor desempenho que outros modelos mais complexos (Webb, 2010). Porém, como esse tipo de abordagem não se aplica aos demais modelos, não devemos levar em consideração apenas estes resultados no julgamento entre a eficácia do modelo proposto e os demais modelos.

Outro ponto bastante relevante sobre os resultados apresentados está relacionado à transparência de quais características serão utilizadas no modelo de aprendizagem de máquina. Este é um tema bastante discutido pelo profissionais de saúde ao utilizarem modelos preditivos no sistema de apoio à decisão médica. Pois, alguns modelos de aprendizagem de máquina não deixam transparente quais propriedades foram avaliadas para se chegar ao resultado preditivo em análise, isso gera uma certa desconfiança por parte destes profissionais, que julgam ser muito importante o perfeito entendimento de todas as propriedades que foram levados em consideração nas predições, pois ao se tratar de vidas, deve-se ter um enorme cuidado e atenção com estes aspectos. Pois, a partir do momento que tem-se uma clareza sobre os fatores utilizados pelos modelos de IA, estes podem ser utilizados pelos profissionais da saúde em sistemas de apoio à decisão médica, fazendo com que seja feita uma análise sobre o paciente através de uma inteligência aumentada. Inteligência aumentada é um termo que define a aplicação de inteligência artificial no processo de tomada de decisão humana em diversas áreas. Na área de cuidado à saúde, este termo está associado à aplicação de IA, por exemplo, no suporte ao diagnóstico ou prognóstico dos pacientes. Neste sentido, a inteligência artificial juntamente com os profissionais de cuidado à saúde vêm desempenhando um papel importante na vida de muitas pessoas (Bazoukis et al., 2022). Quanto maior a transparência sobre os aspectos usados

nos modelos de aprendizagem de máquina, maior poderá ser a confiança dos profissionais da saúde sobre a utilização desta tipo de ferramenta nos cuidados da saúde dos pacientes.

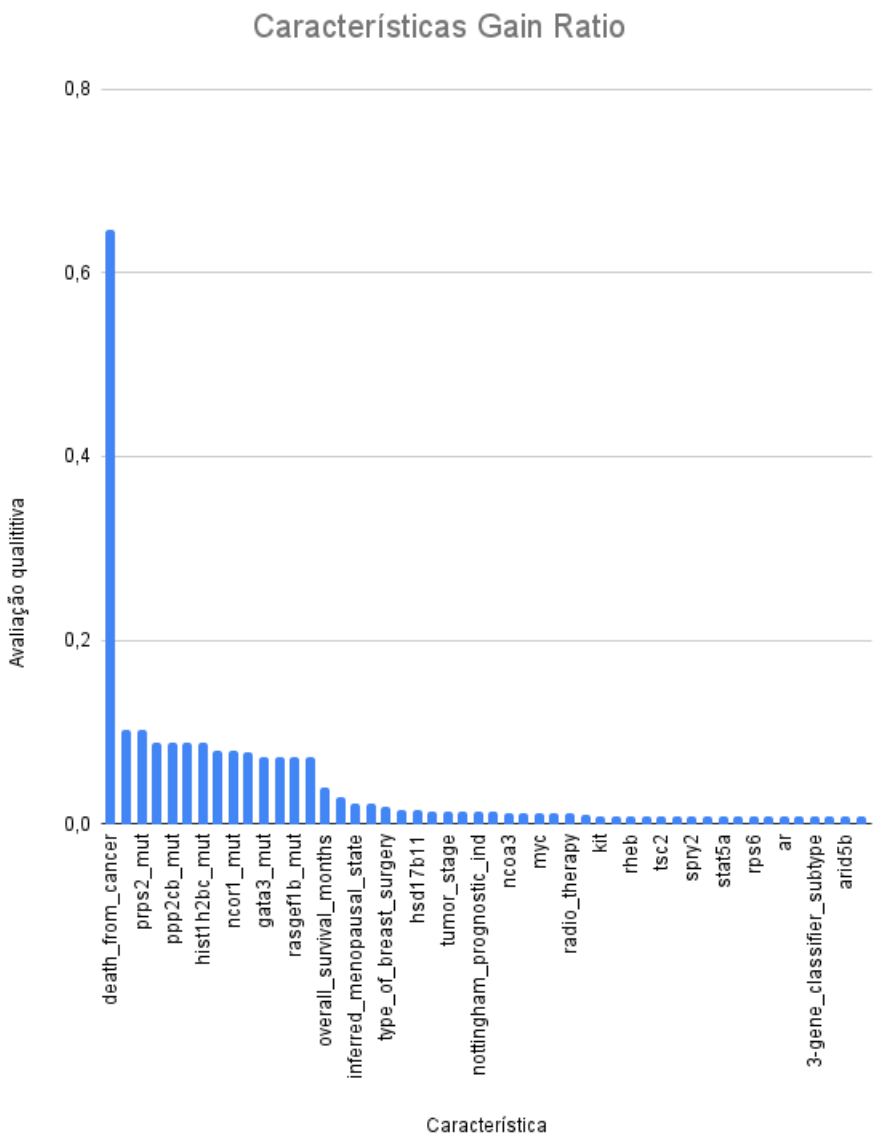

Figura 5.2: Representação das 25 melhores features segundo o algoritmo Gain Ratio para o conjunto de dados do experimento

figures/features_Relief-F.png

Figura 5.3: Representação das 25 melhores features segundo o algoritmo Relief-F para o conjunto de dados do experimento

figures/features_gr_Relief-F.png

Figura 5.4: Representação das 25 melhores features segundo o algoritmo Gain Ratio-Relief-F para o conjunto de dados do experimento

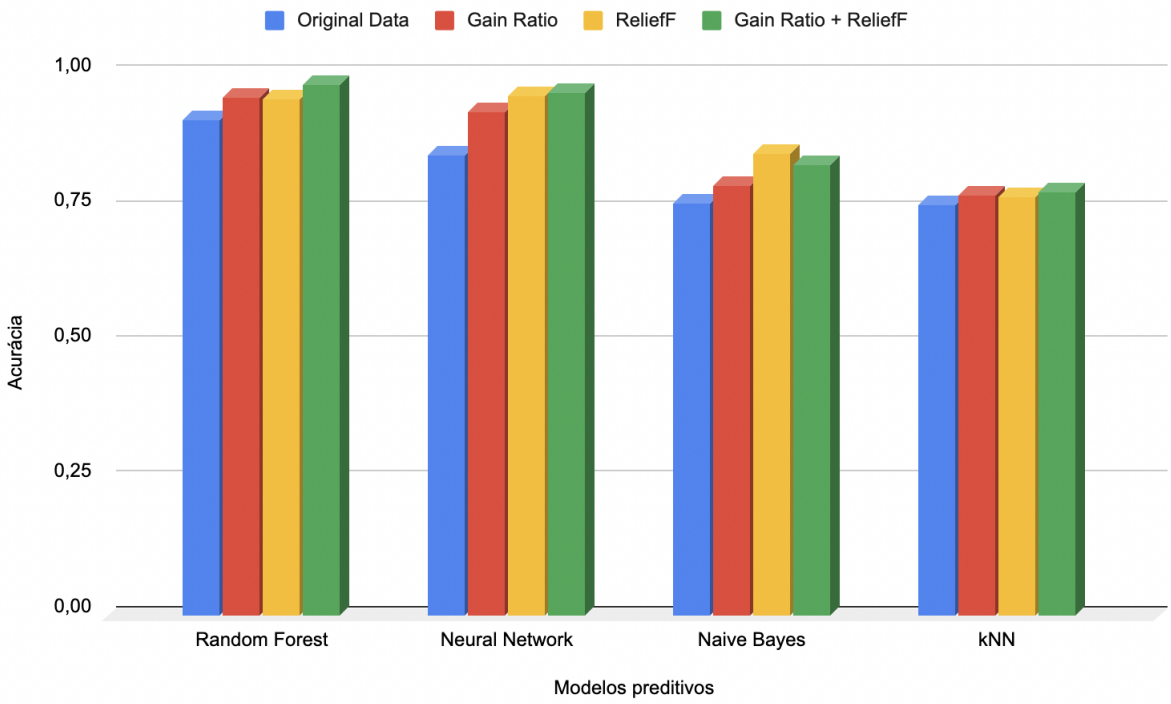

Figura 5.5: Resultados de avaliação da acurácia de cada modelo de acordo com as diferentes abordagens de seleção de características

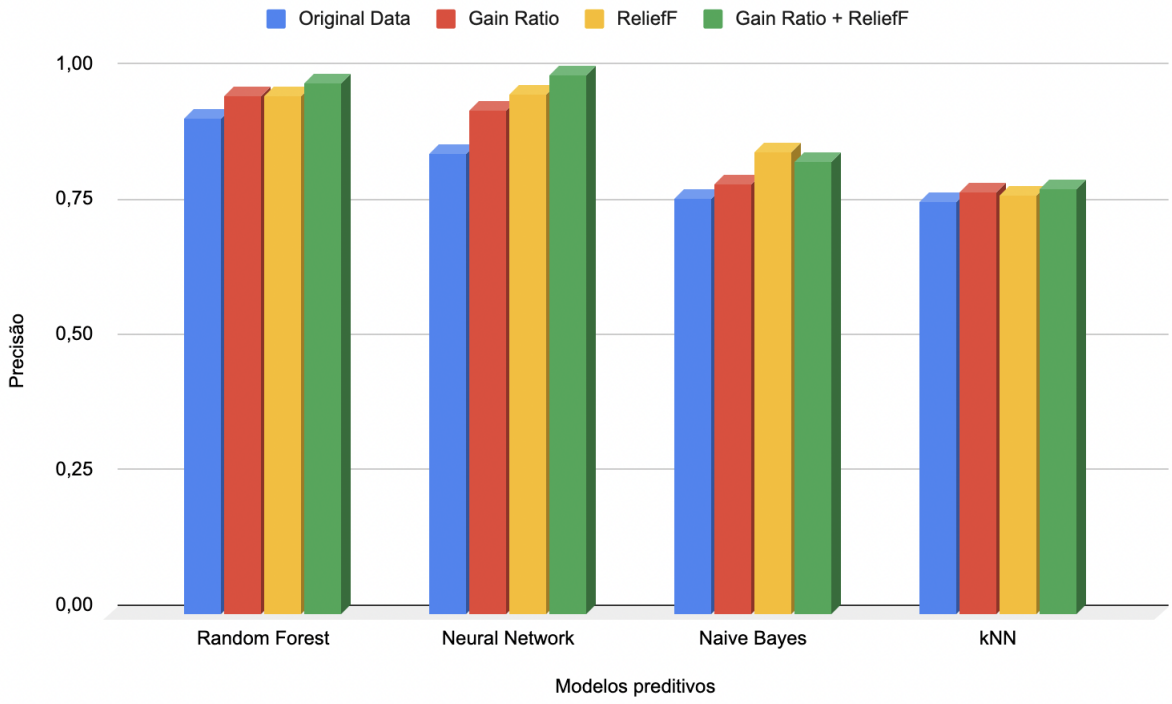

Figura 5.6: Resultados de avaliação da precisão de cada modelo de acordo com as diferentes abordagens de seleção de características

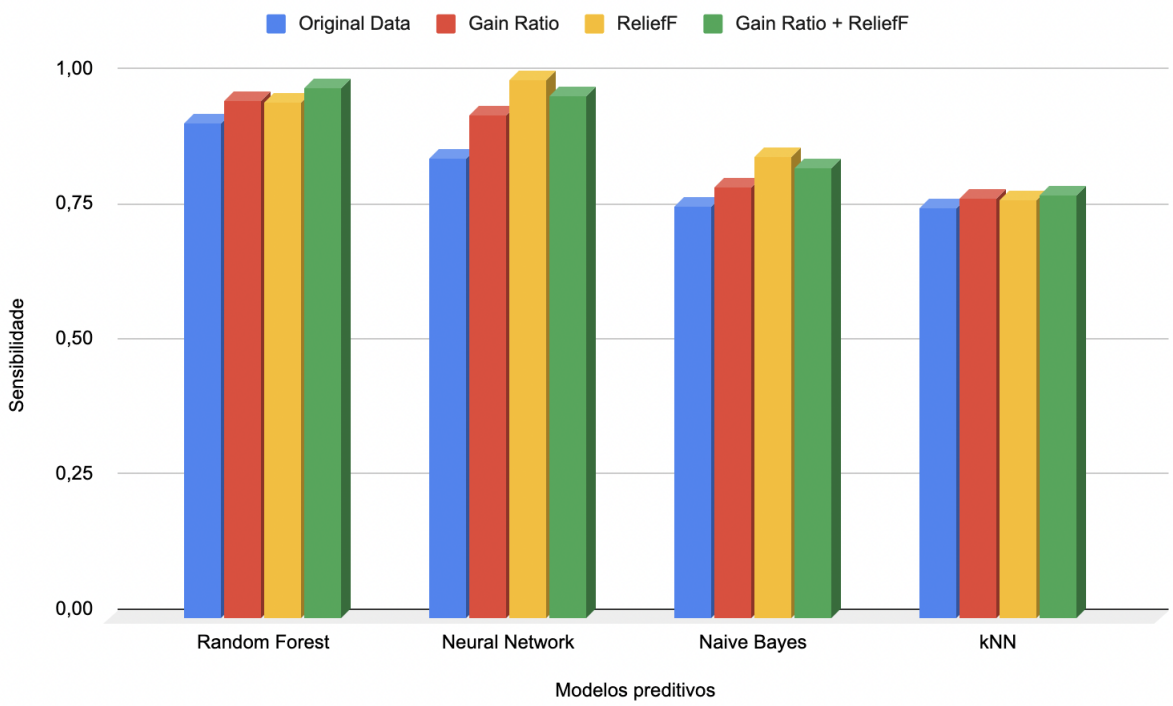

Figura 5.7: Resultados de avaliação de sensibilidade de cada modelo de acordo com as diferentes abordagens de seleção de características

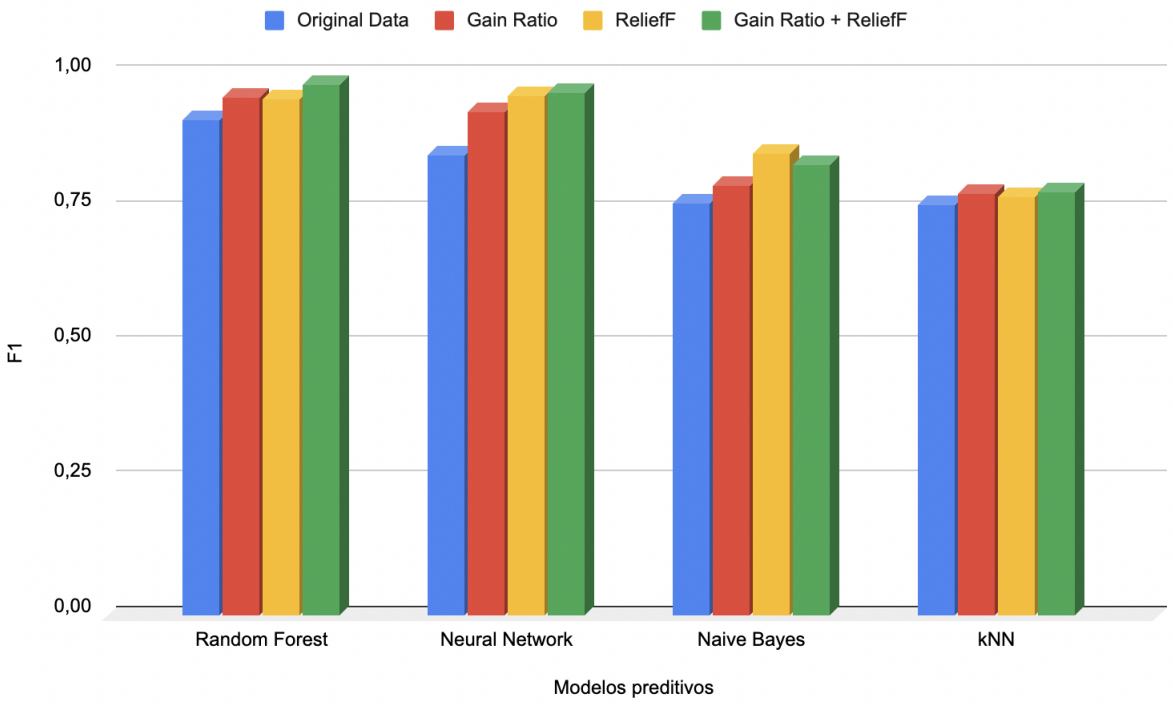

Figura 5.8: Resultados de avaliação F1 de cada modelo de acordo com as diferentes abordagens de seleção de características

6

Considerações Finais

Neste trabalho, apresentamos uma revisão da literatura feita para encontrar quais são os algoritmos de seleção de características mais utilizados na literatura no contexto de prognóstico do câncer de mama, o resultado dos achados científicos foram utilizados com base para definição do objeto de estudo desenvolvido no trabalho aqui proposto. Através da revisão sistemática realizada foram encontrados que dois dos algoritmos mais usando no contexto anteriormente mencionado são: Gain Ratio e RelieF. Além disso, também foi percebido que durantes os últimos anos vem crescendo bastante a utilização de um tipo específico de conjunto de dados no prognóstico da neoplasia de câncer de mama, o conjunto de dados denominamo microarranjo, que é composto de informações genéticas dos pacientes, devidos aos avanços computacionais na área médica, hoje é possível obter esse tipo de dados com maior facilidade, e os resultados de predições obtidos através da utilização deste tipo de conjunto de dados vem se mostrando mais satisfatórios do que quando é utilizado apenas dados clínicos dos pacientes, este tipo de conjunto de dados é característico ter um número muito alto de características, que são os genes, e um número limitado de instâncias. Do ponto de vista médico, é um desafio enorme avaliar um número muito genes e identificar quais deles podem trazer alguma informação relevante sobre o quadro prognóstico do paciente. Portanto, a utilização de um SAD para obter tais variáveis é fundamental para o médico conseguir obter um número reduzido destas variáveis, e consequentemente conseguir dar um prognóstico mais assertivo para o paciente em análise. Após a realização de um estudo prévio para identificação dos pontos fortes e fracos de cada um dos modelos, Gain Ratio e ReliefF, quando aplicados de forma isolada na seleção de características de um conjunto de dados com uma alta dimensão de elementos, que é o caso de microarranjos. Foi definido que seria construído um novo algoritmo, este por sua vez como sendo um algoritmo híbrido, que é a junção de dois os mais modelos já existentes na literatura, afim de obter melhores resultados em relação aos modelos anteriormente mencionados. Sendo assim, foi feito um experimento para implementar o modelo proposto, esta implementação se deu através da utilização do framework Orange, que é bastante utilizado na literatura para implementação de

modelos de aprendizagem de máquina. Para validação dos resultados obtidos através do experimento realizado, foram utilizadas algumas das métricas de validação de modelos mais usadas na literatura, são elas: acurácia, precisão, sensibilidade e F1. Além disso, os modelos utilizados para obter as predições usadas no experimento foram: Random Forest, Neural Network, kNN e Naive Bayes.

Portanto, vale destacar que o modelo implementado foi capaz de trazer melhores resultados para as métricas de avaliação dos modelos utilizados no experimento deste trabalho. Além disso, ele é capaz de diminuir o impacto dos pontos negativos de cada modelo de seleção de características, se utilizados de forma separada para tal finalidade. Sendo assim, os resultados encontrados foram bastante satisfatórios, impactando positivamente na predição do quadro clínico de um paciente com câncer, com melhores resultados na predição deste quadro, o profissional de saúde pode ter uma atenção mais assertiva acerca da evolução do câncer sobre um determinado paciente.

Sobre os aspectos das limitações do trabalho aqui apresentado, podemos relatar a validação dos genes obtidos através da aplicação do modelo seleção de características do algoritmo proposto. Outro fator relevante é não utilização do algoritmo em um cenário real para o contexto da neoplasia de câncer de mama no Brasil, apesar do conjunto de dados ser validado e publicado em uma revista científica de enorme conceito no meio cientifico, ainda se faz necessário a aplicação e validação do mesmo algoritmo em pacientes no cenário nacional, e consequentemente validação dos resultados por um profissional da saúde.

Como trabalhos futuros, fica a possibilidade de construção de um sistema de apoio à decisão médica no contexto do prognóstico do câncer de mama, utilizando o modelo proposto e que o este sistema possa ser utilizado e validado por um profissional de saúde, bem como a utilização do mesmo sobre outros conjuntos de dados para pacientes diagnosticados com câncer de mama.

Os artefatos do projeto aqui desenvolvido podem ser encontrados no seguinte repositório na plataforma GitHub: (Silva, 2022).

Referências

Haneen Banjar, David Adelson, Fred Brown, and Naeem Chaudhri. Intelligent techniques using molecular data analysis in leukaemia: an opportunity for personalized medicine support system. *BioMed Research International*, 2017, 2017.

George Bazoukis, Jennifer Hall, Joseph Loscalzo, Elliott Marshall Antman, Valentín Fuster, and Antonis A Armoundas. The inclusion of augmented intelligence in medicine: A framework for successful implementation. *Cell Reports Medicine*, 3(1):100485, 2022.

BedfordBreastCancer. 3d mammography. URL https://www.bedfordbreastcenter.com/mammogram-los-angeles/.

Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

Joseph D Bronzino. *Biomedical Engineering Handbook 2*, volume 2. Springer Science & Business Media, 2000.

Jie Cai, Jiawei Luo, Shulin Wang, and Sheng Yang. Feature selection in machine learning: A new perspective. *Neurocomputing*, 300:70–79, 2018. ISSN 0925-2312. **DOI** https://doi.org/10.1016/j.neucom.2017.11.077. URL https://www.sciencedirect.com/science/article/pii/S0925231218302911.

Luciano Fernandes Chala and Nestor de Barros. AvaliaÃ§Ãdas mamas com mÃde imagem. *Radiologia Brasileira*, 40:4 – 6, 02 2007. ISSN 0100-3984. URL http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0100-39842007000100001&nrm=iso.

Girish Chandrashekar and Ferat Sahin. A survey on feature selection methods. *Computers Electrical Engineering*, 40(1):16–28, 2014. ISSN 0045-7906. **DOI** https://doi.org/10.1016/j.compeleceng.2013.11.024. URL https://www.sciencedirect.com/science/article/pii/S0045790613003066. 40th-year commemorative issue.

Gary M. Clark, Susan G. Hilsenbeck, Peter M. Ravdin, Michele De Laurentiis, and C. Kent Osborne. Prognostic factors: Rationale and methods of analysis and integration. *Breast*

Cancer Research and Treatment, 32(1):105–112, Jan 1994. ISSN 1573-7217. **DOI** 10.1007/BF00666211. URL https://doi.org/10.1007/BF00666211.

Christina Curtis, Sohrab P Shah, Suet-Feung Chin, Gulisa Turashvili, Oscar M Rueda, Mark J Dunning, Doug Speed, Andy G Lynch, Shamith Samarajiwa, Yinyin Yuan, et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, 486(7403):346–352, 2012.

Janez Demšar, Tomaž Curk, Aleš Erjavec, Črt Gorup, Tomaž Hočevar, Mitar Milutinovič, Martin Možina, Matija Polajnar, Marko Toplak, Anže Starič, Miha Štajdohar, Lan Umek, Lan Žagar, Jure Žbontar, Marinka Žitnik, and Blaž Zupan. Orange: Data mining toolbox in python. *Journal of Machine Learning Research*, 14:2349–2353, 2013. URL http://jmlr.org/papers/v14/demsar13a.html.

P. Domingos and M. Pazzani. On the optimality of the simple Bayesian classifier under zero-one loss. *Machine Learning*, 29:103–130, 1993.

Ruth M. Dunne, Ailbhe C. O'Neill, and Clare M. Tempany. Chapter 9 - imaging tools in clinical research: Focus on imaging technologies. In David Robertson and Gordon H. Williams, editors, *Clinical and Translational Science (Second Edition)*, pages 157–179. Academic Press, second edition edition, 2017. ISBN 978-0-12-802101-9. **DOI** https://doi.org/10.1016/B978-0-12-802101-9.00009-0. URL https://www.sciencedirect.com/science/article/pii/B9780128021019000090.

Mansour Ebrahimi. Comparing various attributes of prolactin hormones in different species: Application of bioinformatics tools. *Iranian Journal of Veterinary Research*, 12, 01 2011.

Khaled Fawagreh, Mohamed Medhat Gaber, and Eyad Elyan. Random forests: from early developments to recent advancements. *Systems Science & Control Engineering: An Open Access Journal*, 2(1):602–609, 2014.

Patrizia Ferroni, Fabio M Zanzotto, Silvia Riondino, Noemi Scarpato, Fiorella Guadagni, and Mario Roselli. Breast cancer prognosis using a machine learning approach. *Cancers*, 11(3): 328, 2019.

AndrÃGonÃ de Freitas, ClÃ¡udio Kemp, Maria Helena Louveira, Sandra Maria Fujiwara, and Leandro Ferracini Campos. Mamografia digital: perspectiva atual e aplicaÃ§Ãfuturas. *Radiologia Brasileira*, 39:287 – 296, 08 2006. ISSN 0100-3984. URL http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0100-39842006000400012&nrm=iso.

Ruffo de Freitas JÃ, Rodrigo Disconzi Nunes, Edesio Martins, Maria Paula Curado, Nilceana Maya Aires Freitas, Leonardo Ribeiro Soares, and JosÃ© Carlos Oliveira. Prognostic factors and overall survival of breast cancer in the city of Goiania, Brazil: a population-based study.

Revista do ColÃBrasileiro de CirurgiÃ, 44:435 – 443, 10 2017. ISSN 0100-6991. URL http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0100-69912017000500435&nrm=iso.

Taís Freire Galvão, Thais de Souza Andrade Pansani, and David Harrad. Principais itens para relatar revisões sistemáticas e meta-análises: A recomendação prisma. *Epidemiologia e serviços de saúde*, 24:335–342, 2015.

Leila Posenato Garcia. Revisão sistemática da literatura e integridade na pesquisa. *Epidemiologia e Serviços de Saúde*, 23:7–8, 2014.

Maximiliano Ribeiro et al. Guerra. jul 2015. **DOI** 10.1590/0102-311X00145214. URL https://doi.org/10.1590/0102-311X00145214.

D. Hand, H. Mannila, and P. Smyth. *Principles of Data Mining*. Cambridge, Massachussetts, USA: MIT Press, 2002.

INCA. Câncer de mama. URL https://www.inca.gov.br/tipos-de-cancer/cancer-de-mama.

Fei Jiang, Yong Jiang, Hui Zhi, Yi Dong, Hao Li, Sufeng Ma, Yilong Wang, Qiang Dong, Haipeng Shen, and Yongjun Wang. Artificial intelligence in healthcare: past, present and future. *Stroke and Vascular Neurology*, 2(4):230–243, 2017. ISSN 2059-8688. **DOI** 10.1136/svn-2017-000101. URL https://svn.bmj.com/content/2/4/230.

A. Karegowda, A. Manjunath, and M. Jayaram. Comparative study of attribute selection using gain ratio and correlation based feature selection. 2010a.

Asha Gowda Karegowda, AS Manjunath, and MA Jayaram. Comparative study of attribute selection using gain ratio and correlation based feature selection. *International Journal of Information Technology and Knowledge Management*, 2(2):271–277, 2010b.

Youness Khourdifi and Mohamed Bahaj. Feature selection with fast correlation-based filter for breast cancer prediction and classification using machine learning algorithms. In *2018 International Symposium on Advanced Electrical and Communication Technologies (ISAECT)*, pages 1–6. IEEE, 2018.

Barbara Kitchenham and Stuart Charters. Guidelines for performing systematic literature reviews in software engineering. 2007.

Konstantina Kourou, Themis P Exarchos, Konstantinos P Exarchos, Michalis V Karamouzis, and Dimitrios I Fotiadis. Machine learning applications in cancer prognosis and prediction. *Computational and structural biotechnology journal*, 13:8–17, 2015.

Rajnish Kumar, Anju Sharma, and Rajesh Kumar Tiwari. Application of microarray in breast cancer: An overview. *Journal of pharmacy & bioallied sciences*, 4(1):21, 2012.

Moshe Lichman et al. Uci machine learning repository, 2013.

Batta Mahesh. Machine learning algorithms-a review. *International Journal of Science and Research (IJSR).[Internet]*, 9:381–386, 2020.

Claudia Mazo, Cathriona Kearns, Catherine Mooney, and William M Gallagher. Clinical decision support systems in breast cancer: a systematic review. *Cancers*, 12(2):369, 2020.

David Moher, Douglas G Altman, Alesandro Liberati, and Jennifer Tetzlaff. Prisma statement. *Epidemiology*, 22(1):128, 2011.

Ryoichi Nakahori, Ryuji Takahashi, Momoko Akashi, Kana Tsutsui, Shino Harada, Roka Matsubayashi, Shino Nakagawa, Seiya Momosaki, and Yoshito Akagi. Breast carcinoma originating from a silicone granuloma: A case report. *World Journal of Surgical Oncology*, 13:72, 02 2015. **DOI** 10.1186/s12957-015-0509-6.

PJ Nogueira, JN Martins, B Lemos, E Góis, A Elavai, G Rodrigues, Carlos Matias Dias, L Couceiro, I Alves, A Lourenço, et al. Relatório do grupo de trabalho de estatísticas da saúde. 2012.

Yongjin Park, Stanley Shackney, and Russell Schwartz. Network-based inference of cancer progression from microarray data. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 6(2):200–212, 2008.

Joel S Parker, Michael Mullins, Maggie CU Cheang, Samuel Leung, David Voduc, Tammi Vickery, Sherri Davies, Christiane Fauron, Xiaping He, Zhiyuan Hu, et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. *Journal of clinical oncology*, 27(8): 1160, 2009.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Peter M Ravdin, Laura A Siminoff, Greg J Davis, Mary Beth Mercer, Joan Hewlett, Nancy Gerson, and Helen L Parker. Computer program to assist in making decisions about adjuvant therapy for women with early breast cancer. *Journal of clinical oncology*, 19(4):980–991, 2001.

Payam Refaeilzadeh, Lei Tang, and Huan Liu. Cross-validation. *Encyclopedia of database systems*, 5:532–538, 2009.

Marko Robnik-Šikonja and Igor Kononenko. Theoretical and empirical analysis of relieff and rrelieff. *Machine Learning*, 53(1):23–69, Oct 2003a. ISSN 1573-0565. **DOI** 10.1023/A:1025667309714. URL https://doi.org/10.1023/A:1025667309714.

Marko Robnik-Šikonja and Igor Kononenko. Theoretical and empirical analysis of relieff and rrelieff. *Machine learning*, 53(1):23–69, 2003b.

Claude Sammut and Geoffrey I. Webb. *Encyclopedia of Machine Learning and Data Mining*. Springer Publishing Company, Incorporated, 2nd edition, 2017. ISBN 148997685X.

Arthur L Samuel. Machine learning. *The Technology Review*, 62(1):42–45, 1959.

Cristina Mamédio da Costa Santos, Cibele Andrucioli de Mattos Pimenta, and Moacyr Roberto Cuce Nobre. The pico strategy for the research question construction and evidence search. *Revista latino-americana de enfermagem*, 15:508–511, 2007.

Tim Sherer. Feature selection (data mining) [internet], 2018. https://docs.microsoft.com/en-us/analysis-services/data-mining/feature-selection-data-mining?view=asallproducts-allversions#:~:text=In%20this%20article&text=Feature%20selection%20refers%20to%20the,or%20features0from%20existing%20data/, acessado em 20/04/2021.

Cleyton Cézar Souto Silva, Rodrigo Pinheiro de Toledo Vianna, and Ronei Marcos Moraes. Sistema de apoio a decisão: a segurança alimentar e o modelo em rede neural. *Revista Brasileira de Ciências da Saúde*, 16(1):79–84, 2012.

Maxwell E. A. Silva. Algoritmo híbrido para seleção de características, 2022. https://github.com/maxwellacioli/projeto_ppgi, acessado em 19/08/2022.

Maxwell E. A. Silva, Victor G. L. Holanda, Rodrigo S. Silva, Paulo V. L. Severiano, and Rafael A. Silva. Seleção de características biológicas para prognóstico de câncer: Revisão sistemática da literatura, Dez 2020. URL https://www.jhi-sbis.saude.ws/ojs-jhi/index.php/jhi-sbis/article/view/817.

Maxwell Esdra Acioli Silva, Victor Gabriel Lima Holanda, Rodrigo Santos da Silva, Paulo Victor Laurentino Severiano, and Rafael de Amorim Silva. Seleção de características biológicas para prognóstico de câncer: Revisão sistemática da literatura. *Journal of Health Informatics*, 12, 2021.

Bikesh Kumar Singh. Determining relevant biomarkers for prediction of breast cancer using anthropometric and clinical features: A comparative investigation in machine learning paradigm. *Biocybernetics and Biomedical Engineering*, 39(2):393–409, 2019. ISSN 0208-5216. **DOI** https://doi.org/10.1016/j.bbe.2019.03.001. URL https://www.sciencedirect.com/science/article/pii/S0208521618304261.

Hyuna Sung, Jacques Ferlay, Rebecca L. Siegel, Mathieu Laversanne, Isabelle Soerjomataram, Ahmedin Jemal, and Freddie Bray. Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, n/a(n/a). **DOI** https://doi.org/10.3322/caac.21660. URL https://acsjournals.onlinelibrary.wiley.com/doi/abs/10.3322/caac.21660.

Luiz Antonio Teixeira and Luiz Alves AraÃ Neto. CÃde mama no Brasil: medicina e saÃpÃno sÃXX. *SaÃe Sociedade*, 29, 00 2020. ISSN 0104-1290. URL http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0104-12902020000300313&nrm=iso.

Aaron Theisen. Microarray-based comparative genomic hybridization (acgh). *Nature Education*, 1(1):45, 2008.

Ryan J. Urbanowicz, Melissa Meeker, William La Cava, Randal S. Olson, and Jason H. Moore. Relief-based feature selection: Introduction and review. *Journal of Biomedical Informatics*, 85:189–203, 2018a. ISSN 1532-0464. **DOI** https://doi.org/10.1016/j.jbi.2018.07.014. URL https://www.sciencedirect.com/science/article/pii/S1532046418301400.

Ryan J Urbanowicz, Melissa Meeker, William La Cava, Randal S Olson, and Jason H Moore. Relief-based feature selection: Introduction and review. *Journal of biomedical informatics*, 85:189–203, 2018b.

Paul J Van Diest, Elsken van der Wall, and Jan PA Baak. Prognostic value of proliferation in invasive breast cancer: a review. *Journal of clinical pathology*, 57(7):675–681, 2004.

S Visalakshi and V Radha. A literature review of feature selection techniques and applications: Review of feature selection in data mining. In *2014 IEEE International Conference on Computational Intelligence and Computing Research*, pages 1–6. IEEE, 2014.

World Cancer Research Fund WCRF. Breast cancer statistics. URL https://www.wcrf.org/dietandcancer/cancer-trends/breast-cancer-statistics.

Geoffrey I. Webb. *Naïve Bayes*, pages 713–714. Springer US, Boston, MA, 2010. ISBN 978-0-387-30164-8. **DOI** 10.1007/978-0-387-30164-8$_5$76. URL https://doi.org/10.1007/978-0-387-30164-8_576.

World Health Organization WHO. Cancer. URL https://www.who.int/news-room/fact-sheets/detail/cancer.

Gordon C Wishart, Elizabeth M Azzato, David C Greenberg, Jem Rashbass, Olive Kearins, Gill Lawrence, Carlos Caldas, and Paul DP Pharoah. Predict: a new uk prognostic model that predicts survival following surgery for invasive breast cancer. *Breast Cancer Research*, 12 (1):1–10, 2010.