UNIVERSIDADE DE SÃO PAULO
ESCOLA DE ENGENHARIA DE SÃO CARLOS
FACULDADE DE MEDICINA DE RIBEIRÃO PRETO
INSTITUTO DE QUÍMICA DE SÃO CARLOS

THAINE CRAVO MARQUES DOS SANTOS

# Utilização de aprendizado de máquina para refinamento do grupo de risco de câncer de próstata em pacientes tratados com radioterapia

Ribeirão Preto
2022

THAINE CRAVO MARQUES DOS SANTOS

VERSÃO CORRIGIDA

# Utilização de aprendizado de máquina para refinamento do grupo de risco de câncer de próstata em pacientes tratados com radioterapia

Dissertação apresentada ao Programa de Pós-Graduação Interunidades em Bioengenharia da Escola de Engenharia de São Carlos – Faculdade de Medicina de Ribeirão Preto e Instituto de Química de São Carlos da Universidade de São Paulo, como requisito para a obtenção do Título de Mestre em Ciências.

Orientador: Prof. Dr. Joaquim Cezar Felipe

Ribeirão Preto

2022

Ficha catalográfica elaborada pela Biblioteca Prof. Dr. Sérgio Rodrigues Fontes da EESC/USP com os dados inseridos pelo(a) autor(a).

Santos, Thaine Cravo Marques dos
S237u Utilização de aprendizado de máquina para refinamento do grupo de risco de câncer de próstata em pacientes tratados com radioterapia / Thaine Cravo Marques dos Santos; orientador Joaquim Cezar Felipe. São Carlos, 2022.

Dissertação (Mestrado) - Programa de Pós-Graduação Interunidades em Bioengenharia e Área de Concentração em Bioengenharia -- Escola de Engenharia de São Carlos; Faculdade de Medicina de Ribeirão Preto; Instituto de Química de São Carlos, da Universidade de São Paulo, 2022.

1. Aprendizado de Máquina. 2. Grupo de Risco. 3. Câncer de Próstata. I. Título.

Eduardo Graziosi Silva - CRB - 8/8907

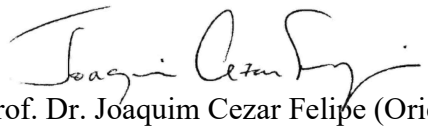
# FOLHA DE JULGAMENTO

**Candidato(a): Thaine Cravo Marques dos Santos**

Título: “Utilização de aprendizado de máquina para refinamento do grupo de risco de câncer de próstata em pacientes tratados com radioterapia”

Data da defesa: 26/09/2022

| **Comissão Julgadora** | **Resultado** |
|---|---|
| Prof(a). Dr(a). Joaquim Cezar Felipe<br>UFSCar - **Orientador** | Não Votante |
| Prof(a). Dr(a). Gustavo Viani Arruda<br>FMRP/USP | Aprovado |
| Prof(a). Dr(a). Marcela Xavier Ribeiro<br>UFSCar | Aprovado |
| Prof(a). Dr(a). Ivan Torres Pisa<br>UNIFESP | Aprovado |



Prof. Dr. Joaquim Cezar Felipe (Orientador)

Presidente da Comissão de Pós-Graduação: Prof. Dr. Adair Roberto Aguiar

# AGRADECIMENTOS

Agradeço aos meus pais por todo suporte e amor.

Agradeço ao meu orientador Prof. Dr. Joaquim Cezar Felipe por todo o tempo de orientação e por todo o suporte desde a minha graduação. Agradeço também ao Prof. Dr. Gustavo Viani Arruda por todo o auxílio durante o estudo.

Agradeço ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) pela bolsa de mestrado no período de abril de 2020 a março 2022.

Por fim, agradeço a Deus por me abençoar e me permitir a realização do mestrado.

# RESUMO

SANTOS, T. C. M. **Utilização de aprendizado de máquina para refinamento do grupo de risco de câncer de próstata em pacientes tratados com radioterapia**. 2022. Dissertação (Mestrado) – Escola de Engenharia de São Carlos, Universidade de São Paulo, São Carlos, 2021.

O câncer de próstata é o segundo câncer mais frequente em homens no mundo. A radioterapia apresenta um papel fundamental no tratamento do câncer de próstata. Para planejamento do tratamento inicial, o paciente é classificado em um grupo de risco, podendo ser baixo, intermediário ou alto risco. Entretanto, esses grupos, normalmente baseados em três dados principais coletados durante o diagnóstico (concentração do antígeno prostático específico pré-terapia, soma do escore de Gleason e estágio clínico do tumor), costumam apresentar alta heterogeneidade e, assim, pacientes com diferentes características podem ser classificados em um mesmo grupo de risco. O tratamento excessivo de tumores com pouca probabilidade de progressão e o subtratamento de tumores mais agressivos são consequências disso. As técnicas de Aprendizado de Máquina vêm sendo utilizadas com sucesso em diversas aplicações voltadas à oncologia, incluindo classificação e avaliação de risco. Com isso, o objetivo deste estudo é refinar a metodologia de determinação do grupo de risco de câncer de próstata em pacientes tratados com radioterapia, utilizando métodos de aprendizado não supervisionado. Foram utilizados dados de 485 pacientes com câncer de próstata tratados com radioterapia entre janeiro de 2010 e janeiro de 2017 no Hospital das Clínicas da Faculdade de Medicina de Ribeirão Preto. Oito diferentes algoritmos de agrupamento foram implementados na linguagem de computação Python, sendo eles: K-means, Mini Batch K-means, Affinity Propagation, Agglomerative Clustering, BIRCH, DBSCAN, OPTICS e Agrupamento Profundo com Autoencoder. Três estratégias foram propostas para obter um melhor resultado do refinamento dos grupos de risco e seus resultados foram analisados e comparados utilizado o método de Kaplan-Meier, juntamente com o teste log-rank, e o método da silhueta. Os algoritmos com os melhores desempenhos foram Mini Batch K-means e Autoencoder, apresentando curvas distintas no gráfico de Kaplan-Meier, os menores valores no teste log-rank e valores positivos de coeficiente de silhueta. Foram implementadas árvores de decisões a partir dos grupos resultantes destes algoritmos para tornar explícitos os indicadores mais relevantes e para apresentar conjuntos de regras que permitem entender a lógica de geração dos agrupamentos. Por fim, uma ferramenta foi desenvolvida para que o usuário consiga classificar novos pacientes nos grupos de risco definidos pelos melhores resultados das três estratégias. Conclui-se que este estudo apresenta novas regras de grupos de risco refinadas, baseadas em

dados referentes ao tumor e ao paciente, além de oferecer uma maneira mais simples e direta para o médico especialista compreender a formação dos grupos de risco.

Palavras-chave: Aprendizado de Máquina; Grupo de Risco; Câncer de Próstata.

# ABSTRACT

SANTOS, T. C. M. **Application of machine learning to refine the prostate cancer risk group in patients treated with radiotherapy**. 2022. Dissertação (Mestrado) – São Carlos School of Engineering, University of São Paulo, 2022.

Prostate cancer is the second most common cancer in men worldwide. Radiotherapy is fundamental for the treatment of prostate cancer. For initial treatment planning, the patient is classified into a risk group, which can be low, intermediate or high risk. However, these groups, usually based on three main data collected during diagnosis (pre-therapy prostate specific antigen concentration, sum of Gleason score and tumor clinical stage), tend to present great heterogeneity and, thus, patients with different characteristics may be classified in the same risk group. Overtreatment of tumors that are unlikely to progress and undertreatment of more aggressive tumors are consequences of this. Machine learning techniques have been used successfully in a variety of oncology applications, including risk assessment and classification. Therefore, the aim of this study is to refine the methodology for determining the risk group for prostate cancer in patients treated with radiotherapy, using unsupervised learning methods. Data from 485 patients with prostate cancer treated with radiotherapy between January 2010 and January 2017 at the Hospital das Clínicas of the Faculty of Medicine of Ribeirão Preto were used. Eight different clustering algorithms were implemented in the Python programming language. The algorithms are: K-means, Mini Batch K-means, Affinity Propagation, Agglomerative Clustering, BIRCH, DBSCAN, OPTICS and Deep Clustering using Autoencoder. Three strategies were proposed to obtain a better result from the refinement of risk groups and their results were analyzed and compared using the Kaplan-Meier method, with the log-rank test, and the silhouette method. The algorithms with the best performances were Mini Batch K-means and Autoencoder, presenting different curves in the Kaplan-Meier graph, the lowest values in the log-rank test and positive silhouette coefficient values. Decision trees were implemented from the groups resulting from these algorithms to make the most relevant indicators explicit and in order to present sets of rules that allow understanding the logic of generating the clusters. Finally, a tool was developed so that the user can classify new patients into the risk groups defined by the best results of the three strategies. It is concluded that this study presents new refined risk group rules, based on tumor and patient data, in addition to offering a friendlier way for the specialist to understand the risk groups.

Keywords: Machine Learning; Risk Group; Prostate Cancer.

# LISTA DE ILUSTRAÇÕES

# LISTA DE TABELAS

# SUMÁRIO

# 1 INTRODUÇÃO

O câncer de próstata é o segundo câncer mais frequente em homens no mundo. No Brasil, estima-se que terão 65840 novos casos para cada ano entre os anos de 2020 e 2022 (INSTITUTO NACIONAL DO CÂNCER, 2020). A radioterapia desempenha um papel fundamental no tratamento do câncer de próstata (SANDA et al., 2018). A classificação em grupos de risco é de grande importância para que o tratamento do paciente seja planejado, pois o grupo de risco é utilizado como base para selecionar as opções apropriadas que devem ser consideradas para o tratamento inicial e para prever a probabilidade de falha bioquímica após a terapia. Uma abordagem amplamente utilizada é aquela que divide os pacientes em três grupos: baixo, intermediário e alto risco. Para esta classificação são utilizados três dados principais coletados durante o diagnóstico: a concentração do antígeno prostático específico (PSA) pré-terapia, a soma do escore de Gleason e o estágio clínico do tumor (D'AMICO et al., 1999; MOHLER et al., 2014).

Entretanto, a divisão nesses três grupos de risco resulta em grupos relativamente heterogêneos, apresentando pacientes com características clínicas distintas em um mesmo grupo (REESE et al., 2012; HERNANDEZ et al., 2007). Essa heterogeneidade pode implicar, para casos superestimados, em tratamentos excessivos de tumores com probabilidade baixa de progressão e, por outro lado, para casos subestimados, em tratamentos insuficientes de pacientes com doença agressiva (BERLIN et al., 2019; ZUMSTEG et al., 2013).

Com isso, em muitas situações, é desejável uma estratificação mais precisa, visando gerar grupos de risco mais acurados (REESE et al., 2012). Para auxiliar nessa tarefa e obter um prognóstico mais preciso, várias ferramentas foram desenvolvidas com base em modelos estatísticos e conceitos de Aprendizado de Máquina (ROSS et al., 2002; SUAREZ-IBARROLA et al., 2019; VAN BOOVEN et al., 2021).

Ao longo dos últimos 20 anos, observa-se o uso de Aprendizado de Máquina em aplicações em câncer com diversos objetivos, como identificar, classificar, detectar ou distinguir tumores. Já o uso em aplicações para previsão e prognóstico do câncer é mais recente e tem como foco a avaliação de risco, a previsão da recorrência do câncer e a previsão da capacidade de sobrevivência ao câncer (CRUZ; WISHART, 2006).

Algoritmos de agrupamento (*clustering)* são algoritmos de aprendizado não supervisionado que agrupam um determinado conjunto de dados em *clusters* com base na similaridade. Assim, os padrões dentro de um mesmo *cluster* (intragrupo) são semelhantes entre

si em comparação ao padrão pertencente a um *cluster* diferente (intergrupo). Esse método é utilizado em diversas análises, como mineração de dados, recuperação de documentos, segmentação de imagens e classificação de padrões (JAIN; MURTY; FLYNN, 1999).

Churilov et al. (2004) utilizaram otimização para aprimorar as regras de agrupamento do grupo de risco de câncer de próstata com o objetivo de obter três grupos de risco mais homogêneos, e Churilov et al. (2005) expandiram o estudo com o método não supervisionado de mapas auto organizáveis para melhorar a previsão da classificação. Gnanapragasam et al. (2016) utilizaram testes estatísticos para criar subgrupos a partir dos grupos de risco baixo, intermediário e alto, a fim de melhorar a predição de morte por câncer. Eminaga et al. (2020) apresentaram um modelo de combinação para classificação do câncer de próstata usando abordagens de Aprendizagem Profunda. Já Zupan et al. (2000) propuseram um framework utilizando técnicas de Aprendizado de Máquina para prever a probabilidade de recorrência do câncer. Diferentes técnicas de Aprendizado de Máquina, incluindo técnicas supervisionadas, também já foram implementadas para classificar pacientes de câncer de próstata e prever a recorrência do câncer.

Diante das possibilidades oferecidas pelo Aprendizado de Máquina, podem surgir alternativas para a classificação dos pacientes com câncer de próstata em grupos de risco, com consequente aperfeiçoamento da qualidade dos grupos, uma vez que é comum que pacientes com características distintas sejam classificados no mesmo grupo na abordagem que classifica em três grupos de risco (baixo, intermediário e alto). Além disso, novas abordagens de Aprendizado de Máquina, incluindo técnicas baseadas em Aprendizado Profundo, também podem melhorar a precisão na estimativa do risco de falha bioquímica e identificar quais indicadores são mais relevantes para realizar o agrupamento dos pacientes.

Portanto, este projeto tem como objetivo buscar um refinamento do grupo de risco de câncer de próstata em pacientes tratados com radioterapia utilizando algoritmos de Aprendizado de Máquina não supervisionados para propor, implementar e avaliar diferentes métodos de agrupamento que possibilitam obter grupos de risco mais homogêneos e melhorar a precisão na estimativa do risco de falha bioquímica. Para isso, diferentes estudos foram realizados e comparados utilizando idade, dados clínicos e laboratoriais de 485 pacientes e oito diferentes algoritmos de Aprendizado de Máquina não supervisionado.

Os resultados obtidos foram grupos de riscos diferentes estatisticamente entre si em todas as abordagens propostas, com destaque para as abordagens utilizando os algoritmos Mini Batch K-means e Autoencoder. Também foi possível observar e analisar as características de cada grupo formado e identificar os indicadores mais relevantes em cada abordagem. Além

disso, foi implementado um outro método, dessa vez “caixa branca” de predição do grupo de risco com os resultados obtidos, podendo auxiliar, assim, a compreensão do médico especialista a respeito das regras intrínsecas a cada abordagem de classificação, auxiliando na tomada de decisão em relação ao planejamento do tratamento do paciente.

Com isso, pretende-se que este trabalho possa contribuir para que o tratamento do paciente com câncer de próstata seja planejado de forma mais eficaz, de acordo com as características do tumor e do próprio paciente. Assim, o paciente receberá um tratamento mais adequado para o seu caso, diminuindo os casos de tratamentos excessivos para tumores com baixa probabilidade de progressão e de tratamentos insuficientes para tumores mais agressivos.

Este trabalho está estruturado da seguinte forma: o capítulo 2 descreve com maiores detalhes os objetivos geral e específicos; o capítulo 3 apresenta alguns trabalhos relevantes para a área abordada neste projeto; o capítulo 4 apresenta os fundamentos teóricos utilizados para a realização do trabalho; o capítulo 5 descreve os materiais e métodos utilizados para obter os resultados finais; o capítulo 6 apresenta os resultados e discussões; o capítulo 7 apresenta a conclusão e os possíveis trabalhos futuros.

# 2 OBJETIVOS

A definição do grupo de risco de câncer de próstata é de suma importância para o planejamento do tratamento do paciente. Se classificado de forma incorreta, o paciente não receberá o tratamento mais adequado, pois poderá receber um tratamento excessivo mesmo se o tumor apresentar baixa probabilidade de progressão ou um tratamento insuficiente para um tumor mais agressivo (BERLIN et al., 2019; ZUMSTEG et al., 2013). Portanto, os grupos devem ser estabelecidos de forma homogênea e bem definidos, sendo distintos entre si. A abordagem mais amplamente utilizada atualmente divide os pacientes em baixo, intermediário e alto risco, com base na análise de três indicadores: a concentração de PSA pré-terapia, a soma do escore de Gleason e o estágio clínico do tumor (D'AMICO et al., 1999). A despeito de ser bastante adotada como referência, esta abordagem pode resultar em grupos relativamente heterogêneos, prejudicando a confiabilidade do processo.

Diante disto, o objetivo geral deste trabalho é obter um refinamento dos grupos de risco em pacientes com câncer de próstata tratados com radioterapia, propondo diferentes abordagens para formação dos grupos. Os agrupamentos são gerados com o uso de técnicas não supervisionadas de Aprendizado de Máquina tradicionais, complementadas com um estudo preliminar utilizando técnicas mais recentes baseadas em Aprendizagem Profunda, denominada de Autoencoder.

Os objetivos específicos são:

- Identificar indicadores relevantes para o melhor refinamento dos grupos;
- Estabelecer diferentes estratégias de obtenção dos agrupamentos e avaliar pontos fortes e fracos de cada uma;
- Comparar os diferentes algoritmos de aprendizado não supervisionado, incluindo métodos baseados em Aprendizagem Profunda;
- Implementar uma ferramenta de predição do grupo de risco baseada nos melhores métodos de agrupamento identificados no estudo, fornecendo também uma visualização de regras baseada em aprendizado supervisionado (árvores de decisão) com os grupos de risco resultantes do estudo.

# 3 TRABALHOS CORRELATOS

Técnicas de Aprendizado de Máquina são utilizadas em diversas aplicações em câncer para diferentes objetivos. De acordo com Cruz e Wishart (2006), o uso de Aprendizado de Máquina para a previsão e o prognóstico do câncer são mais recentes em comparação aos objetivos como identificar, classificar ou detectar tumores. O trabalho de Suarez-Ibarrola et al. (2019) mostra que os estudos de câncer de próstata utilizando técnicas de Aprendizado de Máquina ou Aprendizado Profundo normalmente têm como objetivos desenvolver algoritmos para previsão de escore de Gleason, diagnóstico auxiliado por computador por ressonância magnética e resultados cirúrgicos e previsão de recorrência bioquímica. Os estudos demonstram a superioridade desses métodos sobre os métodos estatísticos tradicionais. Além disso, os autores afirmam que a inclusão contínua de dados clínicos, mais treinamentos dos algoritmos de Aprendizado de Máquina e Aprendizado Profundo, e a generalização dos modelos podem aumentar a precisão da previsão e aprimorar a medicina individualizada.

Segundo Van Booven et al. (2021), o rastreamento do diagnóstico e a estratificação de risco subsequentes de níveis de PSA, biópsias guiadas por ressonância magnética, biomarcadores genômicos e escore de Gleason são, muitas vezes, submetidos a uma subjetividade significativa. Segundo os autores, a inteligência artificial pode auxiliar médicos a gerenciar e reconhecer relacionamentos entre dados que seriam muito difíceis e demorados para realizar apenas com trabalho humano. Portanto, o uso de ferramentas de inteligência artificial pode reduzir a subjetividade citada e também auxiliar na redução no uso de recursos, melhorar a eficiência e a precisão no diagnóstico e no tratamento do câncer de próstata.

Ainda de acordo com Van Booven et al. (2021), essa tecnologia estará entre algumas das ferramentas essenciais para os patologistas urológicos e para o campo da urologia, à medida que continua a melhorar e ajudar no prognóstico do paciente ao longo do tempo. Para demonstrar isso, os autores descrevem diversos estudos que associam e utilizam Inteligência Artificial e Aprendizado de Máquina, incluindo estudos com o objetivo de criar um sistema de classificação para estratificação de risco.

Na literatura, para esses diferentes objetivos que abrangem Aprendizado de Máquina e câncer de próstata, há estudos que utilizam tanto técnicas de aprendizado supervisionado quanto não supervisionado para prever a recidiva bioquímica e refinar os grupos de risco aprimorando as regras de agrupamento ou adicionando subgrupos de risco. A Tabela 1 mostra quais termos

foram utilizados para a busca dos trabalhos apresentados neste capítulo, as bases de dados utilizadas e o resumo das informações de cada trabalho.

Tabela 1 - Resumo dos trabalhos correlatos apresentados neste trabalho.

| Trabalho | Termos de busca | Base de dados | Metodologia utilizada | Principal objetivo |
|---|---|---|---|---|
| Cruz e Wishart (2006) | machine learning prostate cancer | PubMed | Revisão de literatura | Estudos sobre previsão e prognóstico do câncer |
| Suarez-Ibarrola et al. (2019) | | | | Estudos sobre aplicações recentes de aprendizado de máquina e profundo na prática urológica |
| Van Booven et al. (2021) | | | | Analisar os avanços da tecnologia e seu papel atual no diagnóstico e gerenciamento do câncer de próstata |
| Churilov et al. (2004) | machine learning prostate cancer risk group | Google Scholar | Otimização | Agrupar pacientes em três grupos de risco mais homogêneos |
| Churilov et al. (2005) | | | Otimização e mapas auto organizáveis | |
| Eminaga et al. (2020) | | PubMed | Aprendizado de máquina e aprendizado profundo | Classificação de alterações no escore de Gleason e estágio clínico do tumor |
| Zupan et al. (2000) | machine learning prostate cancer risk group recurrence | PubMed | Diferentes algoritmos de aprendizado de máquina | Construir modelos de classificação a partir de dados de sobrevivência |
| Gnanapragasam et al. (2016) | prostate cancer risk group | PubMed | STATA (métodos estatísticos) | Criar subgrupos a fim de melhorar a predição de morte por câncer |
| Arvaniti et al. (2018) | deep learning prostate cancer recurrence | Google Scholar | Rede neural convolucional em imagens histológicas | Classificação automática do escore de Gleason |
| Kumar et al. (2017) | | | Aprendizado profundo em imagens histológicas | Prever a recorrência bioquímica do câncer |
| Shukla et al. (2018) | machine learning cancer risk group | PubMed | Mapas auto organizáveis e DBSCAN | Melhor compreensão da capacidade de sobrevivência do câncer |

Fonte: Elaborado pela autora.

Churilov et al. (2004) implementaram técnicas de agrupamento baseadas em otimização com o objetivo de agrupar pacientes com câncer de próstata em três grupos de risco mais homogêneos. Para o agrupamento, foram utilizados os dados idade do paciente, estágio clínico do tumor, escore de Gleason e o valor de PSA de 258 pacientes tratados entre 1990 e 1997 no *William Buckland Radiotherapy Center* (WBRC), na Austrália. O resultado do agrupamento consistiu de 10 *clusters*, que posteriormente foram aglutinados para formar três grupos de risco, baseados na porcentagem de pacientes livres de recidiva bioquímica em um período de 5 anos.

Os grupos de risco resultantes foram baixo, intermediário e alto, com 71,6%, 53,1% e 35,7% de pacientes livres de recidiva bioquímica, respectivamente. Como conclusão, os autores afirmaram que a abordagem proposta pode apoiar a tomada de decisão clínica, melhorando a precisão da avaliação de risco e pode, portanto, ser vista como uma ferramenta preditiva baseada em evidências com alta capacidade de geração de conhecimento. Entretanto, trata-se de um estudo que analisa somente a porcentagem de indivíduos livres de recidiva bioquímica em 5 anos, e não analisa as características dos indivíduos pertencentes a cada grupo nem o tempo médio de sobrevida de cada grupo. Apesar de 53,1% dos indivíduos do grupo intermediário estarem livres de recidiva bioquímica em 5 anos, não há informações do tempo em que o restante apresentou recidiva bioquímica, por exemplo. Também não há análise sobre o valor das variáveis dos indivíduos de cada grupo ou como essas variáveis podem influenciar na recidiva bioquímica e, consequentemente, na decisão do grupo de risco.

Churilov et al. (2005) expandiram o estudo anterior, utilizando o método não supervisionado de mapas auto organizáveis para melhorar a previsão da classificação. O resultado incluiu uma árvore de decisão com novas regras de agrupamento, com o objetivo de tornar os novos grupos de risco baixo, intermediário e alto mais homogêneos. Segundo os autores, a metodologia proposta possibilita que as regras sejam refinadas e aprimoradas, sem alterá-las radicalmente e de forma irreconhecível. Neste estudo, os autores incluíram informações sobre as variáveis e como seus os valores podem influenciar na tomada de decisão do grupo de risco. Porém, assim como o estudo anterior, também não há informações sobre o tempo livre de recidiva bioquímica, apenas a porcentagem de indivíduos em cada grupo livres de recidiva bioquímica em 5 anos.

Zupan et al. (2000) propuseram um framework utilizando técnicas de Aprendizado de Máquina para construir modelos de classificação a partir de dados de sobrevivência e, assim, prever a probabilidade de recorrência bioquímica do câncer de próstata. Para isso, foram utilizados dados de dois grupos distintos, sendo um grupo com dados pré-operatórios de 967 pacientes e outro grupo com dados pós-operatórios de 996 pacientes, todos com operação entre junho de 1983 e dezembro de 1996. As variáveis consideradas no trabalho foram: valor de Gleason primário, valor de Gleason secundário, estágio clínico do tumor e PSA. As técnicas de Aprendizado de Máquina foram avaliadas e comparadas a um modelo de risco proporcional de Cox, método estatístico criado especificamente para análise de sobrevivência. Os autores observaram que técnicas relativamente simples de Aprendizado de Máquina conseguem ser tão eficientes quanto um modelo estatístico frequentemente usado e que o uso de modelos de Aprendizado de Máquina pode potencialmente apoiar a tomada de decisão do urologista.

Utilizando técnicas de Aprendizado de Máquina supervisionado e de Aprendizado Profundo, Eminaga et al. (2020) apresentaram um modelo de combinação para classificação de alterações no escore de Gleason e estágio clínico do tumor em pacientes com câncer de próstata. Modelos de Aprendizado Profundo e sete abordagens de Aprendizado de Máquina foram comparados no estudo. Foram utilizadas nove variáveis no total (raça, idade no momento do diagnóstico, PSA pré-operatório, estágio clínico do tumor, estágio de nódulo patológico, estágio de metástase clínica, escore de Gleason, número de núcleos positivos na biópsia e o número total de núcleos de biópsia) de 44321 pacientes da base de dados estadunidense SEER (*Surveillance, Epidemiology, and End Results*). Os modelos de Aprendizado Profundo alcançaram maior precisão na classificação do que outras abordagens. A conclusão dos autores foi que o modelo é uma ferramenta útil de auxílio à decisão clínica para o câncer de próstata e para agrupar pacientes com câncer de próstata em grupos clinicamente significativos.

Além de técnicas de Aprendizado de Máquina, há estudos que utilizam métodos estatísticos para realizar o refinamento do grupo de risco. Gnanapragasam et al. (2016) utilizaram testes estatísticos para criar subgrupos a fim de melhorar a predição de morte por câncer de próstata. Primeiramente, os pacientes foram classificados como de baixo, intermediário ou alto risco e, em seguida, foram criados subgrupos dentro de cada categoria de risco. O grupo de risco baixo não foi dividido, pois os subgrupos provenientes desse grupo não apresentaram diferenças significativas. O grupo de risco intermediário foi dividido em dois grupos, assim como o grupo de alto risco. Portanto, o estudo resultou em cinco grupos de risco. Foram utilizados dados de 10139 pacientes com câncer de próstata do Reino Unido e as variáveis estágio clínico do tumor, escore de Gleason, valor inicial de PSA e a pontuação ISUP (classificação determinada pela Sociedade Internacional de Patologia Urológica). Os autores concluíram que o novo modelo tem um bom desempenho e melhora a previsão de mortalidade, assim como também fornece uma melhor distinção de subgrupos de pacientes para auxiliar na tomada de decisão clínica. Cabe observar que neste estudo há duas variáveis contendo a mesma informação. A pontuação ISUP é definida de acordo com o escore de Gleason. Com isso, a variável escore de Gleason poderia ser descartada e ser utilizada somente a pontuação ISUP.

Além de estudos com dados clínicos, também há estudos que utilizam imagens histológicas de câncer de próstata para obter grupos de risco e prever a recorrência bioquímica, especialmente utilizando métodos de Aprendizado Profundo. O estudo de Arvaniti et al. (2018) tem como objetivo principal treinar uma rede neural convolucional para classificação automática do escore de Gleason em imagens de tecido de câncer de próstata. Como etapa final do trabalho, os autores usaram as previsões do modelo para atribuir pacientes em grupos de

baixo, intermediário e alto risco e estudar a estratificação de sobrevivência baseada somente no escore de Gleason. Os estimadores de Kaplan-Meier mostraram diferenças mais significativas para os grupos de risco obtidos a partir da previsão do modelo utilizando Aprendizado Profundo do que aqueles obtidos por especialistas. Já o trabalho de Kumar et al. (2017) utilizou Aprendizado Profundo para prever a recorrência bioquímica do câncer de próstata após a prostatectomia radical a partir de imagens de tecido. O valor de AUC obtido como resultado do estudo foi de 0,81 e, com isso, os autores concluíram que a abordagem apresentada pode servir como uma ferramenta adicional para auxiliar o patologista, o planejamento de tratamento eficaz e a tomada de decisão.

O estudo de análise de sobrevivência e classificação de pacientes utilizando Aprendizado de Máquina também é realizado com dados de pacientes com outros tipos de câncer. Shukla et al. (2018) realizaram um estudo com o objetivo de desenvolver um modelo analítico de dados robusto para auxiliar em uma melhor compreensão da capacidade de sobrevivência do câncer de mama com dados ausentes e também sobre os fatores associados à sobrevivência do paciente. Além disso, o estudo também buscou estabelecer grupos de pacientes que compartilham características semelhantes. Para isso, os autores utilizaram duas técnicas de Aprendizado de Máquina não supervisionado, mapas auto organizáveis e DBSCAN, para agrupar os pacientes. Em seguida, os grupos resultantes com padrões associados foram usados para treinar a técnica de aprendizado supervisionado perceptron multicamadas (MLP) para a previsão de sobrevivência. Os dados utilizados foram de 85189 pacientes da base de dados SEER. O resultado do agrupamento consistiu de nove grupos, com diferentes tempos de sobrevivência. A divisão dos pacientes em grupos melhorou a precisão da previsão de sobrevivência geral com base no MLP. Assim, a conclusão foi que a abordagem totalmente orientada a dados com base em métodos de aprendizagem não supervisionada melhora a compreensão e auxilia a identificação de padrões associados à capacidade de sobrevivência do paciente.

A maioria dos trabalhos relatados utilizaram as variáveis escore de Gleason, valor de PSA e estágio clínico do tumor para realizar as classificações dos grupos de risco ou prever a recidiva bioquímica. No presente estudo, foram utilizadas sete variáveis para refinar os grupos de risco, incluindo variáveis clínicas não relacionadas diretamente com o tumor. Além disso, essas variáveis também foram analisadas para que sejam identificados os indicadores relevantes para o melhor refinamento dos grupos.

Outro ponto em comum nos estudos apresentados é o uso de apenas um método para o refinamento dos grupos de risco ou previsão da recidiva bioquímica. Apenas o trabalho de

Eminaga et al. (2020) comparou diferentes métodos de Aprendizado de Máquina, porém com o objetivo de prever alterações no escore de Gleason e no estágio clínico do tumor. O presente estudo implementou e comparou diferentes métodos de aprendizado não supervisionado, incluindo métodos baseados em Aprendizagem Profunda.

Além disso, este trabalho realizou estudos de diferentes estratégias para obter os grupos de riscos, que foram analisadas e comparadas. Portanto, este estudo contribui para um entendimento mais completo a respeito do uso de diferentes métodos de Aprendizado de Máquina em dados de pacientes com câncer de próstata tratados com radioterapia. Com isso, o estudo também apresenta uma análise mais profunda das características dos pacientes, e não apenas do tumor, que podem influenciar na tomada de decisão do grupo de risco e, consequentemente, na previsão do tempo livre de recidiva bioquímica.

# 4 FUNDAMENTOS TEÓRICOS

Neste capítulo serão abordados os fundamentos teóricos computacionais utilizados neste estudo. Estes fundamentos abrangem conceitos de Aprendizado de Máquina, agrupamento (Aprendizado de Máquina não supervisionado), diferentes classes dos métodos de agrupamento e os algoritmos mais utilizados, assim como agrupamento utilizando Aprendizado Profundo e árvores de decisão.

## 4.1 Agrupamento com Aprendizado de Máquina

Aprendizado de Máquina é uma subárea da inteligência artificial que permite desenvolver um modelo matemático com base em dados de amostra, com o objetivo de prever ou tomar decisões sem ser explicitamente programado para realizar a tarefa. Os métodos de Aprendizado de Máquina utilizam como base resultados anteriores para aprimorar seu desempenho. De forma geral, um modelo de Aprendizado de Máquina apresenta um elemento de aprendizado e um elemento de desempenho. O ambiente fornece informações para o elemento de aprendizagem, que utiliza então essas informações para modificar o elemento de desempenho para melhorar a tomada de decisão (YAO; LIU, 2013; ZHANG, 2020).

As técnicas de Aprendizado de Máquina incluem aprendizado supervisionado e não supervisionado. O aprendizado supervisionado consiste em aprender uma função a partir de um conjunto de exemplos de entrada-saída. É quando há informação sobre as classes verdadeiras às quais os dados pertencem e essas informações já conhecidas são utilizadas para a modelagem e posterior classificação de novos dados. No aprendizado não supervisionado, não há saída específica fornecida e as classes dos dados não são conhecidas (THEODORIDIS; KOUTROUMBAS, 2009; YAO; LIU, 2013).

A técnica de agrupamento (*clustering*) é uma técnica de aprendizado não supervisionado que tem como objetivo agrupar os dados em diferentes subconjuntos (*clusters*). É o processo de particionar um conjunto de objetos em subconjuntos, de forma que cada subconjunto contenha objetos semelhantes e objetos em diferentes subconjuntos sejam diferentes (ACKERMANN et al., 2014). Diferentes métodos de agrupamento podem gerar diferentes agrupamentos no mesmo conjunto de dados. Como é executado por um algoritmo sem interferência humana, o particionamento pode levar à descoberta de grupos anteriormente desconhecidos nos dados (HAN; KAMBER; PEI, 2012).

Os principais métodos de agrupamento podem ser classificados em métodos de particionamento, métodos hierárquicos e métodos baseados em densidade. Os algoritmos de agrupamento podem ser subclassificados em um ou mais métodos, uma vez que um mesmo algoritmo pode implementar diferentes métodos (HAN; KAMBER; PEI, 2012).

### 4.1.1 Métodos de Particionamento

O método fundamental e mais simples é o método de particionamento, que organiza os objetos de um conjunto de dados em vários grupos exclusivos. Os algoritmos que utilizam o método de particionamento geram todos os grupos simultaneamente como uma partição dos dados e não impõem uma estrutura hierárquica. O número de grupos a serem obtidos é o ponto inicial para o agrupamento e estes são formados para otimizar um critério de particionamento, de modo que os objetos pertencentes a um mesmo grupo sejam semelhantes, ou seja, tenham características semelhantes entre si, e diferentes dos objetos que pertencem a outros grupos (HAN; KAMBER; PEI, 2012; JAIN, 2010).

#### 4.1.1.1 K-means

O algoritmo mais popular e que utiliza o método de particionamento é o K-means. Embora tenha sido proposto pela primeira vez há mais de 50 anos, o K-means ainda é um dos algoritmos mais amplamente utilizados para agrupamento devido a sua facilidade de implementação, simplicidade e eficiência (JAIN, 2010).

Seja $X = \{x_i, i = 1, \dots, n\}$, o conjunto de $n$ pontos a serem agrupados em um conjunto de $K$ grupos, $C = \{C_k, k = 1, \dots, K\}$. Técnicas de particionamento baseado em centroide, como o K-means, utiliza o centroide de um grupo, $c_i$, para representar este grupo. O centroide de um grupo é seu ponto central, podendo ser obtido a partir da média dos objetos atribuídos ao grupo, por exemplo.

A diferença entre um objeto $p \in C_i$ e o centroide $c_i$ é medida por $dist\,(p, c_i)$, onde $dist\,(x, y)$ é a distância entre dois pontos $x$ e $y$. Neste trabalho será utilizada a distância euclidiana. A qualidade do grupo $C_i$ pode ser medida pela variação dentro do grupo, que é a soma do erro quadrático entre todos os objetos em $C_i$ e o centroide $c_i$, definida pela Equação 1.

$$E = \sum_{i=1}^{k} \sum_{p \in C_i} dist(p, c_i)^2 \quad (1)$$

Onde $E$ é a soma do erro quadrático para todos os objetos no conjunto de dados, $p$ é o ponto no espaço que representa um determinado objeto e $c_i$ é o centroide do grupo $C_i$. O objetivo do K-means é minimizar a soma do erro quadrático sobre todos os $K$ grupos possíveis (HAN; KAMBER; PEI, 2012; JAIN, 2010).

O algoritmo define o centroide de um grupo como o valor médio dos pontos dentro de um grupo. Primeiramente, $k$ dos objetos em $X$ são selecionados aleatoriamente, cada um dos quais representa inicialmente uma média ou o centro do grupo. Para cada um dos objetos restantes, um objeto é atribuído ao grupo ao qual é mais semelhante, com base na distância entre o objeto e a média do grupo. O algoritmo K-means melhora iterativamente a variação dentro do grupo. Para cada grupo, é calculada a nova média usando os objetos atribuídos ao grupo na iteração anterior. Assim, todos os objetos são atribuídos novamente utilizando as médias atualizadas como os novos centros de grupo. As iterações continuam até que a atribuição esteja estável, ou seja, os grupos formados na iteração atual são iguais aos formados na iteração anterior (HAN; KAMBER; PEI, 2012).

#### 4.1.1.2 Mini Batch K-means

O algoritmo Mini Batch K-means foi proposto por Sculley (2010) como uma alternativa para o K-means com a vantagem de reduzir o custo computacional, pois não utiliza todo o conjunto de dados a cada iteração, mas sim uma subamostra. O algoritmo consiste em utilizar pequenos lotes (*batch*) aleatórios de tamanho fixo como exemplos para serem armazenados na memória. A cada iteração, uma nova amostra aleatória do conjunto de dados é obtida e utilizada para atualizar os grupos. Esse processo é repetido até a convergência. Cada lote atualiza os grupos usando uma combinação convexa dos valores dos protótipos e dos exemplos, aplicando uma taxa de aprendizado que diminui com o número de iterações. Essa taxa de aprendizado é o inverso do número de exemplos atribuídos a um grupo durante o processo. Conforme o número de iterações aumenta, o efeito de novos exemplos é reduzido, então a convergência é obtida quando não ocorrem alterações nos grupos após várias iterações consecutivas (BÉJAR ALONSO, 2013).

#### 4.1.1.3 Affinity Propagation

Assim como os algoritmos K-means e Mini Batch K-means, o Affinity Propagation, proposto por Frey e Dueck (2007), também é um método que utiliza centroides. Entretanto, diferente dos algoritmos anteriores, o Affinity Propagation considera simultaneamente todos os pontos de dados como exemplos em potencial, e não apenas um conjunto dos dados. Cada ponto é visto como um nó em um grafo e mensagens são transmitidas recursivamente ao longo das arestas dos grafos até que seja criado um conjunto de exemplares apropriado. As mensagens são atualizadas com base em fórmulas simples que buscam minimizar uma determinada função de energia escolhida. O algoritmo recebeu o nome de "propagação de afinidade" (*affinity propagation*) pelo fato de que a magnitude de cada mensagem reflete a afinidade (*affinity*) que um determinado ponto tem para escolher outro ponto de dados como seu exemplar.

Frey e Dueck (2007) desenvolveram um algoritmo que recebe como entrada uma coleção de similaridades com valor real entre os pontos de dados, onde a similaridade $s(i,k)$ indica o quão adequado é o item $k$ para ser exemplar do item $i$. Quando o objetivo é minimizar o erro quadrático, cada similaridade é definida como um erro quadrático negativo. Considerando os pontos $x_i$ e $x_k$, temos que $s(i,k) = -||x_i - x_k||^2$. Não é necessário indicar previamente o número de grupos desejado, pois a entrada do algoritmo é um número real $s(k,k)$ para cada ponto $k$, de modo que os pontos com valores maiores que $s(k,k)$ sejam mais propensos a serem escolhidos como exemplares. Esses valores são chamados de preferências. O número de exemplares identificados (ou seja, o número de grupos) é influenciado pelos valores das preferências de entrada.

Há dois tipos de mensagens trocadas entre os pontos de dados: responsabilidade e disponibilidade. A responsabilidade $r(i,k)$ é enviada do ponto de dados $i$ para o ponto $k$ do exemplar candidato e reflete a evidência acumulada de quão adequado o ponto $k$ é para ser exemplar do ponto $i$, levando em consideração outros potenciais exemplares para o ponto $i$. A disponibilidade $a(i,k)$ é enviada do ponto candidato exemplar $k$ ao ponto $i$ e reflete a evidência acumulada de quão apropriado seria para o ponto $i$ escolher o ponto $k$ como exemplar, levando em consideração os outros pontos que podem escolher o ponto $k$ como exemplar. As mensagens podem ser combinadas em qualquer momento para decidir quais pontos são exemplares.

Inicialmente, os valores das mensagens de disponibilidade recebem o valor zero, ou seja, $a(i,k) = 0$, e os valores de responsabilidades são calculadas de acordo com a Equação 3.

$$r(i,k) \leftarrow s(i,k) - \max_{k' \neq k}\{a(i,k') = s(i,k')\} \quad (3)$$

Como os valores das disponibilidades são inicializadas em zero, na primeira iteração $r(i,k)$ é definido como a semelhança de entrada entre o ponto $i$ e o ponto $k$ como seu exemplar menos a maior das semelhanças entre o ponto $i$ e outros candidatos a exemplar.

Nas próximas iterações, quando alguns pontos são atribuídos a outros exemplares, suas disponibilidades assumem valores negativos de acordo com a regra de atualização que pode ser observada na Equação 4.

$$a(i,k) \leftarrow min\left\{0, r(k,k) + \sum_{i' \notin \{i,k\}} max\{0, r(i',k)\}\right\} \quad (4)$$

Os valores negativos das disponibilidades diminuem os valores de algumas similaridades $s(i,k')$ e assim removem os candidatos a exemplares da competição. Essa atualização permite que todos os candidatos exemplares concorram por um ponto e também permite que as evidências de um ponto ser um bom exemplar sejam reunidas.

A disponibilidade $a(i,k)$ é a auto responsabilidade $r(k,k)$ mais a soma das responsabilidades positivas de candidatos a exemplares que $k$ recebe de outros itens. Para limitar a influência de uma responsabilidade positiva muito alta, a soma não pode ser maior que zero. A atualização da auto responsabilidade $a(k,k)$ é feita de forma diferente (Equação 5).

$$a(k,k) \leftarrow \sum_{i' \neq k} \max\{0, r(i',k)\} \quad (5)$$

Esta atualização reflete a evidência acumulada de que o ponto $k$ é um exemplar, com base nas responsabilidades positivas enviadas ao candidato exemplar $k$ de outros pontos.

As regras de atualização requerem apenas cálculos simples que podem ser implementados facilmente e as mensagens precisam ser trocadas apenas entre pares de pontos com semelhanças conhecidas. As disponibilidades e responsabilidades podem ser combinadas para identificar exemplares em qualquer momento durante a propagação de afinidade. Para o ponto $i$, o valor de $k$ que maximiza $a(i,k) + r(i,k)$ também identifica o ponto $i$ como um exemplar se $k = i$, ou identifica o ponto que é o exemplar do ponto $i$. A troca de mensagens

pode ser terminada após um determinado número de iterações, após as mudanças nas mensagens ficarem abaixo de um limiar ou após as decisões locais permanecerem constantes por um determinado número de iterações.

### 4.1.2 Métodos Hierárquicos

Os métodos hierárquicos, diferentemente dos métodos de particionamento, realizam o agrupamento dos objetos em uma hierarquia de grupos. Um método hierárquico pode ser classificado como aglomerativo, em que a hierarquia é formada de "baixo para cima" (abordagem *bottom-up*) ou divisivo, em que a hierarquia é formada de "cima para baixo" (abordagem *top-down*) (HAN; KAMBER; PEI, 2012).

No método hierárquico aglomerativo, inicialmente cada objeto representa um grupo próprio e os grupos são unidos sucessivamente até que a estrutura desejada seja obtida ou até que seja formado um único grupo. Já no método hierárquico divisivo, inicialmente todos os objetos pertencem a um único grupo e, em seguida, esse grupo é dividido em subgrupos sucessivamente, até que a estrutura de grupos desejada seja obtida (ROKACH; MAIMON, 2005). Nos dois métodos, o número desejado de grupos pode ser informado como uma condição de terminação (HAN; KAMBER; PEI, 2012). Como resultado dos métodos hierárquicos, um dendrograma é formado, representando o agrupamento aninhado de padrões e níveis de similaridade. O dendrograma pode ser "quebrado" em diferentes níveis para produzir diferentes agrupamentos de dados e para obter o nível de similaridade desejado (JAIN; MURTY; FLYNN, 1999).

#### 4.1.2.1 Agglomerative Clustering

O Agglomerative Clustering é um algoritmo que utiliza o método hierárquico aglomerativo e é um dos mais amplamente utilizados entre os métodos de agrupamento (SASIREKHA; BABY, 2013).

Para determinar quais grupos serão combinados, o algoritmo encontra dois grupos mais próximos de acordo com uma medida de similaridade e combina-os para formar um único grupo. A escolha dessa métrica tem influência na formação dos grupos, pois a distância entre dois objetos pode ser menor ou maior dependendo da métrica de distância escolhida.

As medidas de ligação são utilizadas para definir os critérios para o cálculo das distâncias entre os objetos em um algoritmo aglomerativo. As mais utilizadas são: distância mínima (Equação 6), distância máxima (Equação 7) e distância média (Equação 8), onde $|p - p'|$ é a distância entre dois objetos $p$ e $p'$; $m_i$ é a média do grupo $C_i$; e $n_i$ é o número de objetos em $C_i$ (HAN; KAMBER; PEI, 2012).

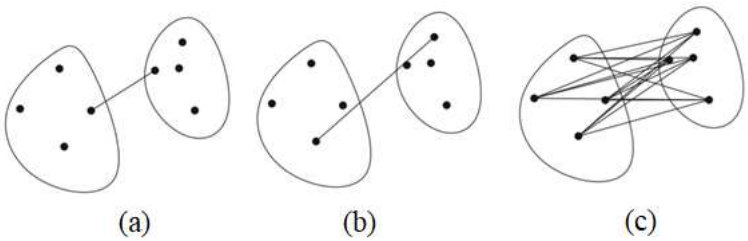
$$dist_{min}(Ci, Cj) = \min_{p \in C_i, p' \in C_j} \{|p - p'|\} \tag{6}$$

$$dist_{max}(Ci, Cj) = \max_{p \in C_i, p' \in C_j} \{|p - p'|\} \tag{7}$$

$$dist_{avg}(Ci, Cj) = \frac{1}{n_i n_j} \sum_{p \in C_i, p' \in C_j} \{|p - p'|\} \tag{8}$$

O algoritmo que utiliza a função de distância mínima é chamado de algoritmo de ligação única (*single linkage*) e combina os grupos mais próximos, que apresentam a menor distância mínima. O algoritmo que implementa a função de distância máxima é chamado de algoritmo de ligação completa (*complete-linkage*) e utiliza as distâncias máximas entre dois conjuntos. O algoritmo que utiliza a distância média (*average-linkage*) é uma solução para melhorar o problema de sensibilidade a *outliers* dos algoritmos anteriores (Figura 1) (HAN; KAMBER; PEI, 2012).

Figura 1 - Exemplos de agrupamento: (a) algoritmo de ligação única, (b) algoritmo de ligação completa, (c) algoritmo de ligação média




Fonte: Adaptado de Nielsen (2016).

Outra ligação também utilizada em algoritmos aglomerativos, chamada de ligação Ward (*Ward linkage*), proposta por Ward (1963), visa combinar dois grupos em um grupo em que a variância seja mínima (NIELSEN, 2016).

#### 4.1.2.2 BIRCH

O algoritmo BIRCH (*Balanced Iterative Reducing and Clustering Hierarchies*), proposto por Charikar et al. (1997), é um método hierárquico aglomerativo adequado para grandes bancos de dados. Dada uma quantidade limitada de memória principal, o algoritmo minimiza o tempo necessário para operações de entrada/saída (E/S) (HAN; KAMBER; PEI, 2012).

O BIRCH agrupa de forma incremental e dinâmica os pontos de dados métricos multidimensionais de entrada para tentar produzir o agrupamento de melhor qualidade com os recursos disponíveis, como memória e tempo. O algoritmo pode encontrar um bom grupo com uma única varredura dos dados e melhorar a qualidade do agrupamento com varreduras adicionais. Além disso, também é o primeiro algoritmo de agrupamento proposto na área de banco de dados para lidar de forma eficaz com instâncias que não fazem parte do padrão subjacente (ruídos) (RAFSANJANI; VARZANEH; CHUKANLO, 2012).

De acordo com Han, Kamber e Pei (2012), para o processo de agrupamento, o algoritmo BIRCH utiliza dois conceitos: recurso de agrupamento (*clustering feature - CF*) e árvore de recursos de agrupamento (*clustering feature tree - CF tree*). O recurso de agrupamento é um resumo das estatísticas de um determinado grupo. A partir de um recurso de agrupamento, muitas estatísticas úteis de um grupo podem ser extraídas. Uma árvore de recursos de agrupamento é uma árvore com altura balanceada que armazena os recursos de agrupamento para um agrupamento hierárquico. Há duas fases iniciais no processo. Na Fase 1, o banco de dados é verificado para a construção uma árvore de recursos de agrupamento inicial na memória, que pode ser vista como uma compactação multinível dos dados que tenta preservar a estrutura de agrupamento inerente dos dados. Na Fase 2, o BIRCH aplica um algoritmo de agrupamento para agrupar os nós folha da árvore de recursos de agrupamento, que remove grupos esparsos como *outliers* e combina grupos densos em outros maiores.

Para a Fase 1, a árvore de recursos de agrupamento é construída dinamicamente à medida que os objetos são inseridos, sendo assim um método incremental. Após a inserção do novo objeto, as informações sobre o objeto são passadas para a raiz da árvore. Se o tamanho da memória necessária para armazenar a árvore de recursos de agrupamento for maior do que o tamanho da memória principal, um valor limite pode ser especificado e a árvore é reconstruída.

O processo de reconstrução é executado construindo uma nova árvore a partir dos nós de folha da árvore antiga. Assim, o processo de reconstrução da árvore é feito sem a necessidade de reler todos os objetos. Além disso, uma vez que a árvore de recursos de agrupamento é

construída, qualquer algoritmo de agrupamento, como um algoritmo de particionamento, pode ser implementado com a árvore de recursos de agrupamento na Fase 2.

### 4.1.3 Métodos baseados em densidade

Os métodos hierárquicos e de particionamento apresentam dificuldades em encontrar grupos de forma arbitrária. Para isso, os métodos de agrupamento baseados em densidade modelam grupos como regiões densas no espaço de dados, de forma não convexa, com o objetivo de identificar os grupos e seus parâmetros de distribuição (HAN; KAMBER; PEI, 2012; ROKACH; MAIMON, 2005).

#### 4.1.3.1 DBSCAN

O DBSCAN (*Density-Based Spatial Clustering of Applications with Noise*), proposto por Ester et al. (1996), é um algoritmo que implementa o método baseado em densidade e encontra objetos centrais, ou seja, objetos que possuem vizinhanças densas. Foi projetado para agrupar dados de formas arbitrárias na presença de ruído em bancos de dados espaciais e não espaciais de alta dimensão.

A principal ideia do algoritmo é que para cada objeto de um grupo, a vizinhança de um determinado raio ($Eps$) deve conter um número mínimo de objetos ($MinPts$). Ambos podem ser definidos pelo usuário. A $Eps$-vizinhança (*Eps-neighborhood*) de um ponto arbitrário $p$ é definida como $N_{Eps} = \{q \in D \mid dist(p,q) < Eps\}$, onde $D$ é o banco de dados de objetos. Se os $Eps$-vizinhos de um ponto $p$ contiverem pelo menos um número mínimo de pontos, então esse ponto é chamado de ponto central. O ponto central é definido como $N_{Eps}(P) > MinPts$ (KHAN et al., 2014).

O $Eps$-vizinho de um objeto $p$ é o espaço dentro de um raio e centrado em $p$. Devido ao tamanho fixo da vizinhança definida por $Eps$, a densidade de uma vizinhança pode ser medida simplesmente pelo número de objetos na vizinhança. O valor de $MinPts$ é utilizado para determinar se uma vizinhança é densa ou não, onde as regiões densas são agrupamentos.

Considerando um objeto central $q$ e um objeto $p$, $p$ é diretamente alcançável por densidade de $q$ se $p$ estiver dentro da $Eps$-vizinhança de $q$. Um objeto $p$ é diretamente alcançável por densidade de outro objeto $q$ se, e somente se, $q$ for um objeto central e $p$ estiver no $Eps$-vizinhança de $q$. Usando a relação diretamente alcançável por densidade, um objeto

central pode “trazer” todos os objetos de sua $Eps$-vizinhança para uma região densa (HAN; KAMBER; PEI, 2012).

De acordo com Han, Kamber e Pei (2012), o algoritmo DBSCAN procura os grupos verificando a $Eps$-vizinhança de cada objeto no conjunto de dados. Inicialmente, todos os objetos em um determinado conjunto de dados $D$ são marcados como "não visitados". O DBSCAN seleciona aleatoriamente um objeto não visitado $p$, marca $p$ como “visitado” e verifica se a $Eps$-vizinhança de $p$ contém pelo menos $MinPts$ objetos. Se sim, um novo cluster $C$ é criado para $p$, e todos os objetos na $Eps$-vizinhança de $p$ são adicionados a um conjunto candidato, $N$. Caso contrário, $p$ é marcado como um ponto de ruído. O DBSCAN adiciona iterativamente a $C$ aqueles objetos em $N$ que não pertencem a nenhum grupo. Nesse processo, para um objeto $p'$ em $N$ que carrega o rótulo “não visitado”, o DBSCAN o marca como “visitado” e verifica sua $Eps$-vizinhança. Se a $Eps$-vizinhança de $p'$ tem pelo menos $MinPts$ objetos, esses objetos na $Eps$-vizinhança de $p'$ são adicionados a $N$. DBSCAN continua adicionando objetos a $C$ até que $C$ não possa mais ser expandido, ou seja, até que $N$ esteja vazio. Neste momento, o cluster $C$ está concluído e, portanto, é formado.

Para encontrar o próximo grupo, o DBSCAN seleciona aleatoriamente um objeto não visitado entre os restantes. O processo de agrupamento continua até que todos os objetos sejam visitados.

#### 4.1.3.2 OPTICS

O algoritmo OPTICS (*Ordering Points To Identify the Clustering Structure*), proposto por Ankerst et al. (1999), é um método de agrupamento baseado em densidade, assim como o DBSCAN, que cria uma ordem aumentada do banco de dados que representa sua estrutura de agrupamento baseada em densidade. Diferente do DBSCAN, em que é necessário que o valor de raio máximo de uma vizinhança ($Eps$) e o número mínimo de pontos necessários na vizinhança de um objeto central ($MinPts$) sejam informados, o OPTICS não produz explicitamente um agrupamento de conjunto de dados. O algoritmo produz uma ordem de grupo e não apresenta a dificuldade de utilizar um conjunto de parâmetros globais na análise de agrupamento.

A ordem de grupo obtida pelo OPTICS é uma lista linear de todos os objetos em análise e representa a estrutura de agrupamento baseada em densidade dos dados. Os objetos em um grupo mais denso são listados mais próximos uns dos outros na ordem do grupo. Essa ordenação

é equivalente ao agrupamento baseado em densidade obtido a partir de uma ampla gama de configurações de parâmetros sem exigir que o usuário determine um limite de densidade específico. A ordenação do grupo pode ser usada para extrair informações básicas de agrupamento, como centros de agrupamento ou grupos de forma arbitrária, obter a estrutura de agrupamento intrínseca e fornecer uma visualização do agrupamento (HAN; KAMBER; PEI, 2012).

Para formar os diferentes agrupamentos simultaneamente, os objetos são processados em uma ordem específica. Esta ordem seleciona um objeto que é alcançável por densidade em relação ao valor de $Eps$ mais baixo para que os grupos com densidade mais alta sejam finalizados primeiro. Para este processo, são necessárias as informações de distância do núcleo (*core-distance*) e de distância de alcançabilidade (*reachability-distance*) dos objetos. A distância do núcleo de um objeto $p$ é o menor valor $Eps'$ tal que a vizinhança de $Eps'$ de $p$ tem pelo menos $MinPts$ objetos. Ou seja, $Eps'$ é o limite mínimo de distância que torna $p$ um objeto central. Se $p$ não for um objeto central de acordo com $Eps'$ e $MinPts$, a distância central de $p$ é indefinida. Já a distância de alcançabilidade do objeto $q$ ao objeto $p$ é o valor mínimo do raio que torna a densidade $p$ alcançável a partir de $q$. De acordo com a definição de densidade-alcançabilidade, $q$ deve ser um objeto central e $p$ deve estar na vizinhança de $q$. Portanto, a distância de alcançabilidade de $q$ a $p$ é definida por: $max\{distância\ do\ núcleo\ (q), dist(p,q)\}$. Se $q$ não é um objeto central de acordo com $Eps$ e $MinPts$, a distância de alcançabilidade de $q$ até $p$ é indefinida.

O algoritmo OPTICS calcula uma ordem de todos os objetos em um determinado banco de dados e, para cada objeto, armazena a distância central e uma distância de alcançabilidade adequada. É criada uma lista chamada *OrderSeeds* para gerar a ordem de saída. Os objetos em *OrderSeeds* são classificados pela distância de alcance de seus respectivos objetos centrais mais próximos, ou seja, pela menor distância de alcance de cada objeto.

O processo começa com um objeto arbitrário de entrada como o objeto atual $p$. Ele recupera a $Eps$-vizinhança de $p$, determina a distância do núcleo e define a distância de alcançabilidade como indefinida. O objeto atual $p$ é então gravado na saída.

Se $p$ não for um objeto central, OPTICS simplesmente passa para o próximo objeto na lista *OrderSeeds*. Se $p$ for um objeto central, então para cada objeto $q$, na vizinhança de $p$, OPTICS atualiza sua distância de alcançabilidade de $p$ e insere $q$ em *OrderSeeds* se $q$ ainda não foi processado. A iteração continua até que a entrada seja totalmente consumida e *OrderSeeds* esteja vazio.

A ordenação do grupo de um conjunto de dados pode ser representada graficamente, o que auxilia na visualização e na compreensão da estrutura de agrupamento em um conjunto de dados (HAN; KAMBER; PEI, 2012).

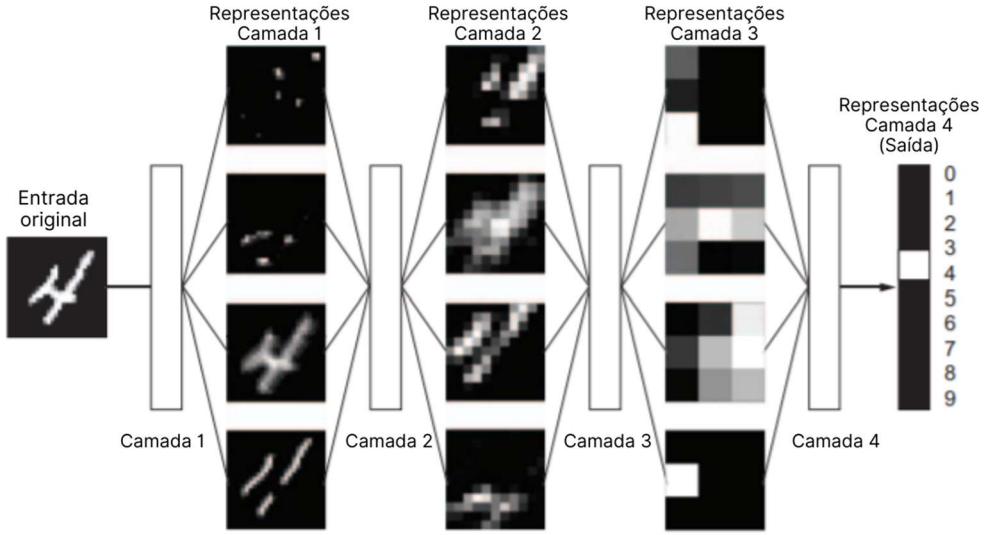
## 4.2 Agrupamento com Aprendizado Profundo

O Aprendizado Profundo é uma subárea específica do Aprendizado de Máquina. Consiste em uma estrutura matemática para aprender representações a partir de dados, em que essas representações são organizadas em camadas. Diferente do Aprendizado de Máquina convencional, o Aprendizado Profundo utiliza uma cascata de camadas de unidades de processamento não linear para extração e transformação de recursos. A primeira camada e a última camada são chamadas, respectivamente, de camada de entrada e camada de saída. As camadas intermediárias entre as camadas de entrada e saída são denominadas de camadas ocultas. As entradas dos algoritmos de Aprendizagem Profunda são transformadas por meio de várias camadas ocultas e as saídas são derivadas do cálculo das camadas ocultas. A quantidade de camadas que contribuem para um modelo de dados é chamada de profundidade do modelo (CHOLLET, 2018; HAO; ZHANG; MA, 2016).

Normalmente, o Aprendizado Profundo envolve dezenas ou até centenas de camadas sucessivas de representações e todas são aprendidas automaticamente com a exposição aos dados de treinamento. Essas representações em camadas geralmente são aprendidas por meio de modelos chamados redes neurais, estruturadas em camadas empilhadas (CHOLLET, 2018). Comparado ao modelo raso, o modelo profundo tem uma melhor capacidade de representação de funções não lineares (HAO; ZHANG; MA, 2016).

Um exemplo de uma rede com múltiplas camadas pode ser visto na Figura 2. No exemplo, a rede transforma a imagem de um dígito para reconhecer qual dígito está representado na imagem. Para isso, a rede transforma a imagem do dígito em representações cada vez mais diferentes da imagem original e cada vez mais informativas para o resultado final.

Figura 2 - Exemplo de uma rede com quatro camadas para reconhecer o dígito 4 que está representado na figura de entrada

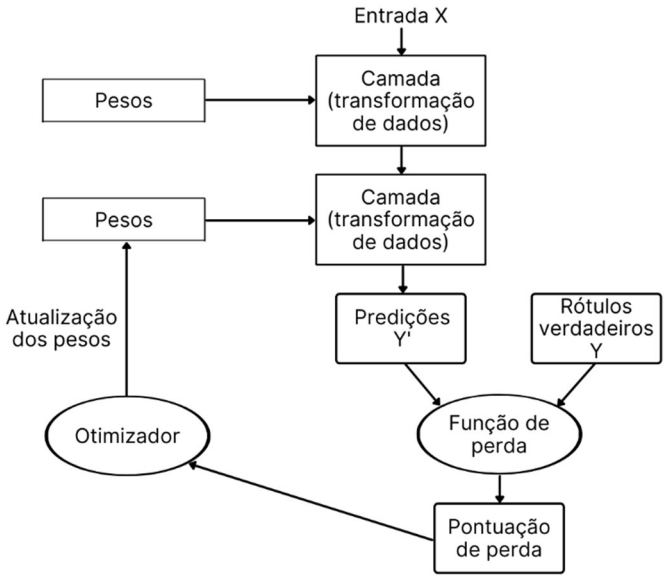



Fonte: Adaptado de Chollet (2018).

A transformação implementada por uma camada é parametrizada por seus pesos. No Aprendizado Profundo, aprender significa encontrar um conjunto de valores para os pesos de todas as camadas de uma rede, de modo que a rede mapeie corretamente as entradas de exemplo para seus rótulos correspondentes.

Para controlar a saída de uma rede neural, é necessário medir o quão distante a saída da rede está do desejado. A função de perda é responsável por realizar essa tarefa, calculando uma pontuação de distância entre a saída da rede e a saída desejada. Essa pontuação é utilizada como um *feedback* para ajustar os valores dos pesos a fim de minimizar a pontuação de perda. Esse processo é realizado pelo otimizador, que implementa o algoritmo de retropropagação, o algoritmo central do Aprendizado Profundo (Figura 3).

Figura 3 - Exemplo de uma rede neural profunda e suas funções




Fonte: Adaptado de Chollet (2018).

Inicialmente, valores aleatórios são atribuídos aos pesos da rede e, com isso, a pontuação de perda é muito alta. A cada exemplo processado pela rede, os pesos são ajustados e a pontuação de perda diminui. Esse processo é chamado de treinamento. Após repetidos exemplos processados, os valores de pesos obtidos minimizam a função de perda. Uma rede com perda mínima em que as saídas são o mais próximo possível do desejado é chamada de rede treinada (CHOLLET, 2018).

De acordo com Min et al. (2018), com o Aprendizado Profundo, as redes neurais profundas podem ser usadas para transformar os dados em representações mais "amigáveis" para o processo de agrupamento, devido à sua propriedade de transformação altamente não linear. O conceito envolvendo Aprendizado Profundo em tarefas de agrupamento de dados é chamado de Agrupamento Profundo (*Deep Clustering*).

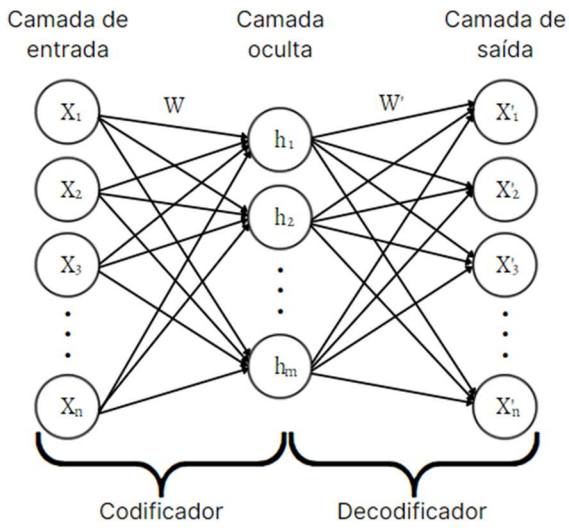
Os métodos tradicionais de agrupamento podem ser classificados como métodos de particionamento, hierárquico e baseados em densidade, como apresentado anteriormente. Já os métodos de Agrupamento Profundo são classificados de acordo com a arquitetura de rede utilizada para o agrupamento, pois o objetivo do Agrupamento Profundo é aprender uma representação orientada a agrupamento. Codificador automático (*Autoencoder*), rede neural de agrupamento, *Generative Adversarial Network* (GAN) e *Variational Autoencoder* (VAE) são exemplos de classificações de Agrupamento Profundo. Além de executar tarefas de agrupamento, os métodos podem gerar novas amostras dos grupos obtidos (CHOLLET, 2018; MIN et al., 2018).

Para orientar as redes a aprender representações amigáveis, há dois tipos de funções de perda de agrupamento: perda de agrupamento principal e perda de agrupamento auxiliar. A perda de agrupamento principal contém os centroides de agrupamento e atribuições de agrupamento de amostras. Ou seja, após o treinamento da rede orientado pela perda de agrupamento, os grupos podem ser obtidos diretamente. A perda de agrupamento auxiliar apenas orienta a rede para aprender uma representação mais viável para o agrupamento, mas não pode gerar grupos de maneira direta. Ou seja, os métodos de Agrupamento Profundo com perda de agrupamento auxiliar requerem a execução de um método de agrupamento após o treinamento da rede para obter os grupos (MIN et al., 2018).

### 4.2.1 Autoencoder

O Autoencoder, também chamado decodificador automático, é um dos algoritmos mais significativos na aprendizagem de representação não supervisionada. O algoritmo obtém um espaço de recursos viável, realizando a compressão de dados. Uma rede Autoencoder fornece uma função de mapeamento não linear através da aprendizagem de um codificador e um decodificador, as duas principais partes que compõem o algoritmo. O codificador é uma função de mapeamento a ser treinada que compacta os dados de entrada para recursos de menor dimensão, e o decodificador deve ser capaz de reconstruir os dados originais a partir dos recursos gerados pelo codificador. Na função de mapeamento, o método garante o mínimo de erro de reconstrução entre a camada do codificador e a camada de dados (MIN et al., 2018).

Figura 4 – Arquitetura básica de um modelo de Autoencoder




Fonte: Adaptado de Zhang, Liu e Jin (2020).

No algoritmo, o modelo de Autoencoder implementado é conectado e simétrico, ou seja, os dados são compactados e descompactados de maneira exatamente oposta. A função recebe uma lista do número de unidades em cada camada do codificador (dims) e o número de camadas do modelo é calculado como $2 * tamanho\ da\ lista\ dims - 1$. Nesta primeira etapa, a camada oculta extrai os recursos, o codificador automático é treinado e os pesos do modelo são salvos para serem utilizados na próxima etapa.

Como a camada oculta geralmente tem dimensionalidade menor do que a camada de dados, ela pode ajudar a encontrar os recursos mais relevantes dos dados. A camada de agrupamento, implementada na próxima etapa do algoritmo, converte a amostra de entrada em um vetor que representa a probabilidade da amostra pertencer a cada grupo. A probabilidade é calculada com a distribuição t-Student. Os parâmetros utilizados nessa etapa são: o número de grupos desejado, a lista com os pesos salvos na etapa anterior e o grau de liberdade na distribuição t-Student. Essa camada age de forma semelhante a um algoritmo de agrupamento, em que os pesos representam os centroides de cada grupo. A última etapa tem como entrada uma variável que contém os dados a serem agrupados e, como saída, o grupo que cada amostra tem maior probabilidade de pertencer.

O algoritmo de Autoencoder, apesar de ser um algoritmo de Aprendizado Profundo, é um algoritmo de aprendizado não supervisionado e, portanto, não necessita de classes previamente definidas para realizar a classificação em diferentes grupos. Com isso, seu desempenho é melhorado não apenas se houver um grande número de instâncias, mas sim se a base de dados for limpa, sem dados duplicados e sem dados faltantes, por exemplo (SALEKSHAHREZAEE; LEEVY; KHOSHGOFTAAR, 2021). Além disso, comparado com outros métodos, algoritmos utilizando Autoencoder podem, em alguns casos, obter uma acurácia razoável em base de dados menores (AMARBAYASGALAN; JARGALSAIKHAN; RYU, 2018).

## 4.3 Árvore de decisão

Árvore de decisão é um método de aprendizado supervisionado não paramétrico de classificação que utiliza a abordagem dividir para conquistar (*divide-and-conquer*). O objetivo é criar um modelo que preveja o valor de uma variável de destino aprendendo regras de decisão inferidas a partir dos atributos dos dados (MYLES et al., 2004; PEDREGOSA et al., 2011).

Para a aprendizagem, o algoritmo de árvore de decisão tem como entrada um conjunto de amostras de treinamento e outro conjunto com seus respectivos rótulos verdadeiros, informando a qual classe cada amostra pertence. Depois de ajustado e treinado, o modelo pode ser utilizado para prever as classes de novas amostras. Há diversas técnicas e algoritmos desenvolvidos para a construção de árvores de decisão. O algoritmo CART (*Classification and Regression Trees*), proposto por Breiman et al. (1984), constrói árvores binárias utilizando o atributo e o limite que geram o maior ganho de informação em cada nó. Também utiliza o índice Gini como medida de impureza para selecionar o atributo, sendo que o valor 0 indica completa igualdade e o valor 1, completa desigualdade. Ou seja, quanto mais próximo de 1, mais amostras de diferentes grupos pertencem ao nó em questão. O atributo com a maior redução de impureza é utilizado para dividir os registros do nó (BRIJAIN et al., 2014; PEDREGOSA et al., 2011).

Sendo $x_i \in R^n$, com $i = 1, \ldots, l$, vetores de treinamento e $y \in R^l$ um vetor de rótulos, a árvore de decisão particiona recursivamente o espaço de recursos de forma que as amostras com os mesmos rótulos ou valores de destino semelhantes sejam agrupadas. Considerando $Q_m$ os dados no nó $m$ com $N_m$ amostras, para cada divisão de candidatos $\theta = (j, t_m)$, com recurso $j$ e limiar $t_m$ , a partição dos dados em $Q_m^{esquerda}(\theta)$ e $Q_m^{direita}(\theta)$ pode ser observada nas Equações 9 e 10.

$$Q_m^{esquerda}(\theta) = \{(x, y) | x_j \leq t_m\} \tag{9}$$

$$Q_m^{direita}(\theta) = Q_m^{esquerda}(\theta)^C \tag{10}$$

Uma vantagem das árvores de decisão é a facilidade de interpretação do modelo construído, pois a visualização da árvore possibilita entender as regras. Com isso, a identificação de atributos importantes e de relacionamentos interclasses pode ser utilizada para apoiar projetos de análises de dados. Outra vantagem é que árvore de decisão utiliza um modelo "caixa branca", em que uma situação observada no modelo pode ter sua condição explicada pela lógica booleana, diferente de uma rede neural artificial que utiliza o modelo "caixa preta" (MYLES et al., 2004; PEDREGOSA et al., 2011).

# 5 MATERIAIS E MÉTODOS

## 5.1 População

Neste estudo retrospectivo, os dados utilizados são de registros de 485 pacientes com câncer de próstata que foram tratados com radioterapia no Hospital das Clínicas da Faculdade de Medicina de Ribeirão Preto (HC-FMRP) entre janeiro de 2010 e janeiro de 2017.

O projeto de pesquisa do presente estudo, juntamente com a solicitação de dispensa de aplicação do Termo de Consentimento Livre e Esclarecido (TCLE), foram submetidos e aprovados pelo Comitê de Ética e Pesquisa do HC-FMRP.

Os critérios de inclusão são: possuir idade entre 44 e 96 anos no momento do diagnóstico e utilização de Radioterapia Conformada 3D (3D-CRT) ou Radioterapia de Intensidade Modulada (IMRT) com uma dose total maior ou igual a 74 Gy.

Os critérios de exclusão são: pacientes com metástases, história prévia de prostatectomia, tratamento quimioterápico ou tratados com radiação pélvica devido ao câncer de próstata e pacientes submetidos à radioterapia hipofracionada.

## 5.2 Dados

A coleta dos dados foi realizada junto ao Serviço de Radioterapia do HC-FMRP, sob a supervisão do Dr. Gustavo Viani Arruda, médico radioterapeuta do Serviço de Radioterapia, no momento do diagnóstico e tratamento dos pacientes.

Todos os pacientes antes do tratamento foram avaliados por seus históricos completos e exame físico. Originalmente, para o planejamento do tratamento, os pacientes foram classificados em grupos de baixo, intermediário e alto risco, de acordo com o método de D'Amico, o qual utiliza o escore de Gleason, o estágio clínico do tumor (estágio T) e o PSA inicial (PSAi) para determinar o grupo.

O escore de Gleason é uma pontuação que avalia o grau histológico do câncer com base no padrão do tumor e é um dos fatores considerados pelo especialista para planejar o tratamento do paciente. Quanto maior o grau, maior a possibilidade de progressão do câncer, com ou sem tratamento. Como o câncer de próstata geralmente é heterogêneo, o escore total de Gleason é composto por duas pontuações: grau primário e grau secundário. O grau primário é referente à maior área do tumor na biópsia, enquanto o grau secundário refere-se à segunda maior área do

tumor. Cada grau varia de 1 a 5. O grau mais baixo 1 indica que as células cancerosas são semelhantes às células normais. Já os maiores graus, como 4 e 5, indicam que as células apresentam uma aparência muito anormal. O escore de Gleason é obtido com a soma destes graus e varia de 2 a 10 (NATIONAL COMPREHENSIVE CANCER NETWORK - NCCN, 2020; NELSON, 2020).

Outro parâmetro utilizado é o estágio T, que representa o tamanho do tumor principal e varia entre T1 e T4. O estágio T1 indica que o tumor não pode ser sentido no exame de toque retal e não pode ser localizado em exames de imagens, porém há a presença de células cancerosas. O tumor em estágio T2 pode ser sentido durante o exame de toque e é encontrado somente na próstata. O estágio T3 indica que o tumor ultrapassou as camadas externas da próstata e pode ser encontrado também na vesícula seminal. O último e mais alto estágio, T4, indica que o tumor se expandiu além da próstata, em estruturas próximas como bexiga, reto e parede pélvica (NCCN, 2020).

O PSA, terceiro parâmetro, é uma proteína produzida por células que revestem pequenas glândulas dentro da próstata. A maioria dos cânceres de próstata tem início nessas células. O nível de PSA é medido a partir de uma amostra de sangue e é o número de nanogramas de PSA por mililitro (ng/mL) de sangue. É utilizado para estabelecer o estadiamento do câncer, planejar e verificar os resultados do tratamento. Após o tratamento, se os níveis de PSA aumentam, é chamado de recidiva bioquímica (ou recidiva do PSA). Isso pode ocorrer devido à volta do câncer (recidiva) ou pelo fato do tratamento não ter reduzido o câncer (persistência) (NCCN, 2020). Entretanto, o PSA não é um marcador específico do câncer de próstata e também é produzido e mensurável em pacientes que não apresentam um tumor ou com doença benigna. Além disso, independente da presença do tumor, o nível de PSA aumenta progressivamente de acordo com a idade e com o volume da próstata (BAYNE; JARRETT, 2016).

Os critérios de classificação utilizados pelo método proposto por D'Amico podem ser observados na Tabela 2. Para o paciente ser classificado no grupo de baixo risco, todos os critérios devem ser cumpridos. Para ser classificado nos grupos intermediário ou alto risco, apenas um dos critérios deve ser cumprido. Todos os pacientes classificados como de alto risco foram submetidos à cintilografia óssea.

Tabela 2 - Critérios D'Amico para a classificação dos grupos de risco

| Grupo de risco | Escore de Gleason | | Estágio clínico do tumor | | Valor de PSAi |
|---|---|---|---|---|---|
| Baixo | < 7 | E | T1 ou T2a | E | < 10 ng/mL |
| Intermediário | 7 | OU | T2b | OU | 10 a 20 ng/mL |
| Alto | > 7 | OU | > T2b | OU | > 20 ng/mL |

Fonte: Elaborado pela autora.

Baseada nas regras do método de D'Amico, um paciente é classificado como baixo risco, se, e apenas se, o escore de Gleason total for menor que 7, o estágio clínico do tumor for T1 ou T2a e o valor de PSAi for menor que 10 ng/mL. Para o paciente ser classificado no grupo de risco intermediário, o valor de escore de Gleason total deve ser 7 ou o estágio clínico do tumor ser T2b ou o valor de PSAi estar entre 10 e 20 ng/mL. Já para classificação no grupo de alto risco, o escore de Gleason total deve ser maior que 7 ou o estágio clínico do tumor maior que T2b ou o valor de PSAi maior que 20 ng/ mL.

No presente estudo, além dos dados escore de Gleason, estágio T e PSAi, também foram coletados outros dados, visando obter uma maior abrangência, assim como um maior entendimento das variáveis que permeiam o posicionamento de um paciente em um determinado grupo de risco.

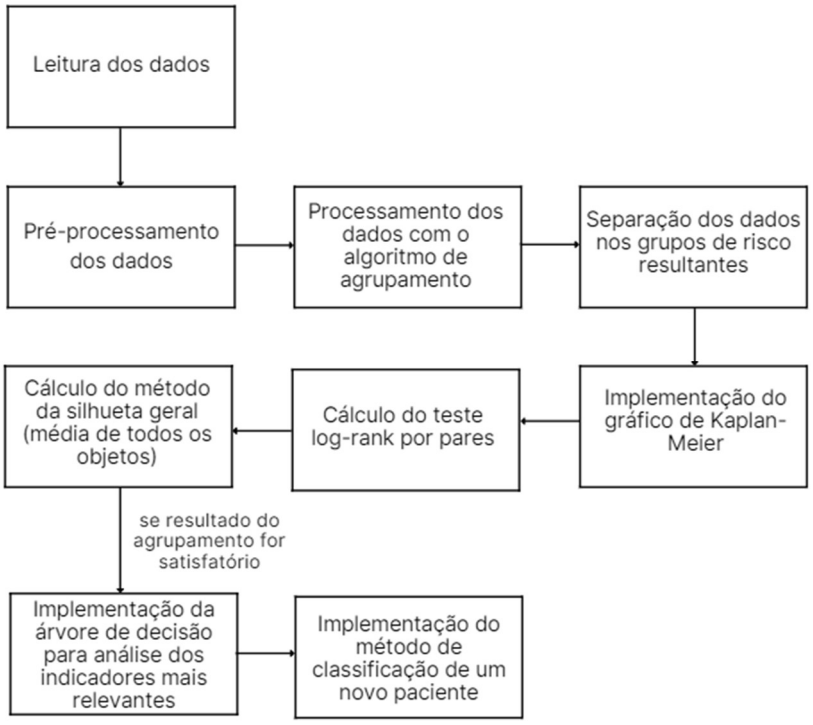
Os seguintes dados adicionais foram coletados:

- idade do paciente no momento do diagnóstico;
- escore de Gleason primário;
- escore de Gleason secundário;
- quantidade de fragmentos retirados para biópsia;
- quantidade de fragmentos positivos;
- ocorrência de invasão perineural do tumor;
- se o paciente tem hipertensão arterial;
- se o paciente tem diabetes;
- ocorrência de recidiva do PSA;
- o tempo de seguimento em meses do paciente.

## 5.3 Metodologia proposta

Para atingir o principal objetivo de refinar os grupos de risco em pacientes com câncer de próstata, foram implementados 8 algoritmos de aprendizado não supervisionado, incluindo o método Autoencoder de Aprendizado Profundo, e os métodos de avaliação de Kaplan-Meier, teste log-rank e método da silhueta. De modo geral, as atividades realizadas podem ser observadas na Figura 5.

Figura 5 - Diagrama com atividades a serem realizadas com os dados




Fonte: Elaborado pela autora.

Inicialmente, os dados de 485 pacientes com todas as variáveis foram lidos para que o pré-processamento fosse realizado. O pré-processamento dos dados inclui a normalização dos mesmos, a verificação se há dados faltantes e a seleção das variáveis utilizadas para o processamento dos dados com os algoritmos de agrupamento. O método utilizado para a normalização foi o PowerTransformer, disponível na biblioteca de pré-processamento de dados na linguagem Python.

Das 12 variáveis coletadas, foram utilizadas 9 para o processamento dos dados: idade do paciente no momento do diagnóstico, valor de PSA inicial, estágio clínico do tumor, escore de Gleason primário, escore de Gleason secundário, porcentagem dos fragmentos positivos da

biópsia (PPB), ocorrência de invasão perineural do tumor, presença de hipertensão arterial do paciente e presença de diabetes. A variável PPB foi criada a partir das variáveis quantidade de fragmentos retirados para biópsia e quantidade de fragmentos positivos, dividindo os valores de quantidade de fragmentos positivos pela quantidade de fragmentos retirados para biópsia e multiplicando o resultado por 100. Para o gráfico de Kaplan-Meier e o teste log-rank, também foram utilizadas as variáveis com as informações de ocorrência de recidiva bioquímica e o tempo de seguimento em meses do paciente.

O processamento dos dados para o refinamento dos grupos de risco foi realizado com a implementação de 8 algoritmos de aprendizado não supervisionado, sendo um deles um método de Aprendizado Profundo. Os algoritmos são: K-means, Mini Batch K-means, Affinity Propagation, Agglomerative Clustering, BIRCH, DBSCAN, OPTICS e Autoencoder. Foram escolhidos algoritmos dos três diferentes métodos de agrupamento, além do método utilizando Aprendizagem Profunda, e os mais citados na literatura de cada método.

Para cada algoritmo, foram obtidos o gráfico de Kaplan-Meier, os valores de teste de log-rank e o valor do coeficiente de silhueta. O gráfico de Kaplan-Meier foi utilizado para que o tempo livre de recidiva bioquímica dos grupos formados por cada algoritmo fosse comparado. Para isso, após o processamento do algoritmo, os dados foram separados de acordo com os grupos resultantes do agrupamento, pois cada grupo gera uma curva no gráfico de Kaplan-Meier. O teste log-rank complementa o gráfico, informando o grau de diferenciação entre os grupos a partir do $p - valor$ calculado. Os valores do teste log-rank foram obtidos por pares, ou seja, todos os grupos foram comparados entre si e foi calculado o $p - valor$ de cada comparação. O método da silhueta foi implementado para analisar se os grupos são compactos, ou seja, se os pacientes pertencentes a um grupo possuem características semelhantes entre si, e se os grupos estão distantes uns dos outros. O resultado do método da silhueta obtido em cada agrupamento é a média dos valores de coeficiente de silhueta dos objetos de todos os grupos, para que o agrupamento completo seja analisado.

A fim de obter um melhor resultado do refinamento dos grupos de risco, 3 estratégias foram elaboradas e estudadas, analisadas e comparadas (Tabela 3):

- **Estratégia 1**: Os algoritmos de agrupamento foram implementados com o objetivo de obter três grupos de risco (baixo, intermediário e alto) a partir das 9 variáveis apresentadas anteriormente. Com os três grupos definidos, as variáveis e as características dos pacientes pertencentes a cada grupo foram analisadas para que as condições de agrupamento fossem definidas.

- **Estratégia 2**: Grupos baixo, intermediário e alto foram gerados de acordo com os critérios propostos por D'Amico, o qual utiliza os valores de escore de Gleason, estágio clínico do tumor e PSA inicial, com as regras descritas na Tabela 2. Esta estratégia visa gerar subgrupos desses, a fim de se obter maior homogeneidade entre os pacientes de cada subgrupo. Primeiramente, apenas os dados dos pacientes do grupo de risco baixo foram utilizados para o agrupamento e, assim, subgrupos do grupo de baixo risco foram definidos. Em seguida, o mesmo processo foi feito apenas com os dados dos pacientes pertencentes ao grupo de risco intermediário e, por último, apenas com os dados dos pacientes do grupo de alto risco. Com isso, os subgrupos de cada grupo de risco foram definidos. O número de subgrupos não foi previamente estabelecido. Após a análise dos resultados, os algoritmos mais adequados para estabelecer os subgrupos foram identificados.
- **Estratégia 3**: Os algoritmos foram utilizados para gerar grupos sem que um limite de número de grupos fosse estabelecido. A cada novo processamento de cada algoritmo de agrupamento, o número de grupos a serem estabelecidos foi aumentado, iniciando por 4 grupos, até que não foram mais identificados novos grupos no gráfico de Kaplan-Meier, ou seja, até que não foram mais identificadas diferenças no tempo livre de recidiva bioquímica entre os novos grupos. Após a análise dos grupos e a comparação dos algoritmos, foram definidos o número de grupos mais adequado e as condições de agrupamento.

Tabela 3 - Resumo das estratégias elaboradas para refinar os grupos de risco

| **Estratégia** | **Número de grupos** | **Principal característica da estratégia** |
|---|---|---|
| 1 | 3 | Alterar as condições de agrupamento |
| 2 | Indefinido | Definir subgrupos a partir dos grupos já definidos |
| 3 | > 3 | Não limitar o número de grupos total |

Fonte: Elaborado pela autora.

Após a implementação das 3 estratégias, os melhores resultados de cada uma foram analisados para definir a estratégia com melhor desempenho, estabelecendo os pontos fortes e fracos de cada uma, o algoritmo de aprendizado não supervisionado que realizou o melhor

agrupamento de acordo com os métodos de avaliação e também os indicadores mais relevantes para obter o refinamento dos grupos de risco.

Para tornar explícitos os indicadores mais relevantes e a fim de apresentar ao usuário conjuntos de regras que permitem a ele entender a lógica de geração dos agrupamentos, foram implementadas árvores de decisão a partir dos indicadores mais relevantes identificados pelos melhores resultados dos algoritmos. O algoritmo de árvore de decisão recebe como parâmetro de entrada o resultado do agrupamento realizado pelos algoritmos de Aprendizado de Máquina, ou seja, os dados de entrada da árvore de decisão é a saída dos algoritmos de agrupamento.

Finalmente, após obter todos os resultados, foi implementada uma ferramenta computacional para predição do grupo de risco a partir dos dados de um determinado paciente, utilizando aprendizado supervisionado com os grupos resultantes. O usuário pode escolher uma ou mais entre as estratégias propostas e depois avaliar os resultados gerados, para escolher a classificação que julgar mais apropriada. Assim, com este trabalho, além de realizar o refinamento do grupo de risco, é possível classificar novos pacientes de acordo com os grupos de risco resultantes do estudo, de forma simples, direta e detalhada.

## 5.4 Ferramentas computacionais

Todos os métodos deste estudo foram implementados na linguagem de programação Python na versão 3.9.12, utilizando o ambiente de desenvolvimento Jupyter Notebook na versão 6.4.8. Os métodos de pré-processamento de dados, os métodos de agrupamento e método da Silhueta foram implementados utilizando a biblioteca de Aprendizado de Máquina scikit-learn na versão 1.0 (PEDREGOSA et al., 2011). O método de Kaplan-Meier e o teste log-rank foram implementados utilizando a biblioteca lifelines (versão 0.26) (DAVIDSON-PILON, 2019) e os gráficos foram implementados com a biblioteca matplotlib (versão 3.4.3) (HUNTER, 2007). As bibliotecas NumPy (versão 1.21.2) (HARRIS et al., 2020), para o processamento de matrizes e com funções matemáticas para realizar operações com matrizes, e pandas (versão 1.3) (McKINNEY, 2010), para análise de dados, também foram utilizadas. Para a implementação do método de agrupamento utilizando Aprendizado Profundo, foi utilizada a biblioteca keras (versão 2.6) (CHOLLET et al., 2015).

Na implementação dos algoritmos de agrupamento, diferentes parâmetros foram utilizados para cada estratégia.

No agrupamento utilizando o K-means, os três parâmetros alterados foram o número de grupos (n_clusters), o número de vezes que o algoritmo roda com diferentes valores de

centroides para obter o melhor resultado (n_init) e o valor para controlar a randomização, para tornar o resultado reprodutível (random_state). Todos os valores foram testados para que fossem obtidos os melhores resultados. Os valores atribuídos a cada parâmetro de acordo com a estratégia podem ser observados na Tabela 4. Os parâmetros não mencionados utilizaram o valor padrão já definido pelo algoritmo.

Tabela 4 – Valores dos parâmetros na implementação do método K-means

| **Estratégia** | | **n_clusters** | **n_init** | **random_state** |
|---|---|---|---|---|
| 1 | | 3 | 50 | 13 |
| 2 | Baixo | 2 | 10 | 75 |
| | Intermediário | 2 | 10 | 19 |
| | Alto | 2 | 10 | 113 |
| 3 | | 4 | 50 | 13 |
| | | 5 | 50 | 42 |

Fonte: Elaborado pela autora.

No algoritmo Mini Batch K-means, os parâmetros alterados foram: número de grupos (n_clusters) e o valor para controlar a randomização, para tornar o resultado reprodutível (random_state). Esses valores em cada estratégia podem ser observados na Tabela 5. Os valores utilizados em outros parâmetros foram os valores padrões já definidos pelo algoritmo.

Tabela 5 – Valores dos parâmetros na implementação do método Mini Batch K-means

| **Estratégia** | | **n_clusters** | **random_state** |
|---|---|---|---|
| 1 | | 3 | 105 |
| 2 | Baixo | 2 | 613 |
| | Intermediário | 2 | 19 |
| | Alto | 2 | 4 |
| 3 | | 4 | 2240 |
| | | 5 | 1170 |

Fonte: Elaborado pela autora.

Já no Affinity Propagation, o único parâmetro definido foi um valor de preferência, que altera o número de grupos definidos pelo algoritmo (preference). Como o algoritmo do Affinity Propagation não contém o parâmetro para definir o número de grupos, esse valor pode ser definido alterando o parâmetro preference. Pontos com valores maiores que o valor de preferência têm maior probabilidade de serem escolhidos como exemplares. Vários valores foram testados para que fosse definido o número de grupos desejado em cada estratégia. Esses

valores podem ser observados na Tabela 6. O restante dos parâmetros foi definido com o valor padrão.

Tabela 6 – Valores do parâmetro preference na implementação do método Affinity Propagation

| **Estratégia** | | **preference** |
|---|---|---|
| 1 | | -250 |
| 2 | Baixo | -100 |
| | Intermediário | -120 |
| | Alto | -160 |
| 3 | 4 grupos | -180 |
| | 5 grupos | -150 |

Fonte: Elaborado pela autora.

Na implementação do Agglomerative Clustering, apenas o parâmetro que define o número de grupos (n_clusters) foi alterado do padrão (Tabela 7).

Tabela 7 – Valores do parâmetro de número de grupos na implementação do método Agglomerative Clustering

| **Estratégia** | | **n_clusters** |
|---|---|---|
| 1 | | 3 |
| 2 | Baixo | 2 |
| | Intermediário | 2 |
| | Alto | 2 |
| 3 | | 4 |
| | | 5 |

Fonte: Elaborado pela autora.

No algoritmo BIRCH, além do parâmetro que define o número de grupos (n_clusters), também foi definido o valor limite do raio do subgrupo obtido pela fusão de uma nova amostra e o subgrupo mais próximo (threshold). Se o raio for maior que o limite definido, um novo subgrupo é criado. Valores baixos promovem a divisão de grupos. Em todas as estratégias, o valor do parâmetro threshold foi de 0.29. Os valores do parâmetro do número de clusters podem ser observados na Tabela 8.

Tabela 8 – Valores do parâmetro de número de grupos e de threshold na implementação do método BIRCH

| **Estratégia** | | **n_clusters** | **threshold** |
|---|---|---|---|
| 1 | | 3 | 0.29 |
| 2 | Baixo | 2 | |

| | | | |
|---|---|---|---|
| | Intermediário | 2 | |
| | Alto | 2 | |
| 3 | | 4 | |
| | | 5 | |

Fonte: Elaborado pela autora.

O algoritmo para realizar o agrupamento com DBSCAN não apresenta o parâmetro de número de grupos. O número de grupos é definido alterando dois parâmetros: o valor da distância máxima entre duas amostras para que uma seja considerada na vizinhança da outra (eps) e o número de amostras em uma vizinhança para que um ponto seja considerado um ponto central (min_samples). Os valores definidos destes parâmetros em cada estratégia para que o número de grupos desejados fosse obtido podem ser observados na Tabela 9.

Tabela 9 – Valores dos parâmetros na implementação do método DBSCAN

| **Estratégia** | | **eps** | **min_samples** |
|---|---|---|---|
| 1 | | 0.385 | 5 |
| 2 | Baixo | 0.4 | 5 |
| | Intermediário | 0.76 | 5 |
| | Alto | 0.8 | 5 |
| 3 | | 0.358 | 4 |
| | | 0.36 | 4 |

Fonte: Elaborado pela autora.

No algoritmo OPTICS, também não há um parâmetro para definir diretamente um número de grupos. Para definir o número de grupos, os parâmetros que definem um o número de amostras em uma vizinhança para um ponto a ser considerado como um ponto central (min_samples) e a distância máxima entre duas amostras para que uma seja considerada na vizinhança da outra (max_eps). Esses valores podem ser observados na Tabela 10. Os outros parâmetros foram definidos de acordo com o padrão do algoritmo.

Tabela 10 – Valores dos parâmetros na implementação do método OPTICS

| **Estratégia** | | **min_samples** | **max_eps** |
|---|---|---|---|
| 1 | | 15 | Não definido |
| 2 | Baixo | 18 | Não definido |
| | Intermediário | 15 | Não definido |
| | Alto | 12 | Não definido |
| 3 | | 12 | Não definido |

| | | |
|---|---|---|
| | 12 | 1.9 |

Fonte: Elaborado pela autora.

A implementação do Autoencoder contém os parâmetros que definem o número de grupos (n_clusters), o valor de corte usado para separar o treinamento em fases distintas (epoch) e o tamanho do conjunto de amostras que será processado em paralelo e que pode aproximar a distribuição dos dados de entrada melhor do que uma única entrada (batch_size), o tipo do otimizador utilizado para treinamento (optimizer) e a função de perda (loss). O número de grupos varia de acordo com a estratégia, assim como mostrado na Tabela 11. Em todas as estratégias, os valores dos parâmetros epoch, batch_size, optimizer e loss utilizados foram os mesmos: 50, 256, adam e mse, respectivamente. Os valores dos parâmetros foram escolhidos de acordo com o mais comumente utilizado em algoritmos de Aprendizado Profundo utilizando codificador automático.

Tabela 11 – Valores do parâmetro de número de grupos na implementação do método Autoencoder

| **Estratégia** | | **n_clusters** |
|---|---|---|
| 1 | | 3 |
| 2 | Baixo | 2 |
| | Intermediário | 2 |
| | Alto | 2 |
| 3 | | 4 |
| | | 5 |

Fonte: Elaborado pela autora.

## 5.4 Validação de agrupamento

Dado um conjunto de dados e um algoritmo de agrupamento, diferentes grupos são obtidos se o algoritmo for executado com diferentes parâmetros. Com isso, é importante que uma avaliação do agrupamento resultante seja feita. A avaliação dos resultados de um algoritmo de agrupamento é conhecida como validação de agrupamento (*cluster validation*).

A validação de agrupamento geralmente se refere à exploração da qualidade de um agrupamento e é mais complexa do que avaliar o ajuste do modelo ou a qualidade da predição na regressão ou classificação supervisionada, pois geralmente as informações sobre as classes verdadeiras dos objetos não estão disponíveis na análise de agrupamento (HENNIG et al., 2016).

Os métodos de validação de agrupamento podem ser divididos em dois tipos: métodos extrínsecos e métodos intrínsecos. Se as informações sobre as classes verdadeiras estiverem disponíveis, um método extrínseco pode ser utilizado para comparar os grupos resultantes do agrupamento obtido pelo algoritmo com as classes verdadeiras. Quando não há informações sobre as classes verdadeiras dos objetos, um método intrínseco pode ser utilizado para avaliar a qualidade do agrupamento, considerando quão distantes estão os grupos resultantes (HAN; KAMBER; PEI, 2012).

### 5.4.1 Método da Silhueta

O método da Silhueta é um método intrínseco proposto por Rousseeuw (1987) que permite que um grupo seja representado por uma silhueta. O método avalia quão distantes estão os grupos e o quão compacto um grupo é.

Considerando um conjunto de dados $D$, de $n$ objetos, suponha que $D$ seja particionado em $k$ grupos. Para cada objeto $o \in D$, é calculado $a(o)$ como a distância média entre $o$ e todos os outros objetos no grupo ao qual $o$ pertence, e $b(o)$ como a distância média mínima de $o$ a todos os grupos aos quais $o$ não pertence. Com isso, o coeficiente de silhueta do objeto $o$ pode ser calculado de acordo com a Equação 11.

$$s(o) = \frac{b(o) - a(o)}{\max\{a(o), b(o)\}} \tag{11}$$

O valor do coeficiente de silhueta está entre -1 e 1. O valor de $a(o)$ reflete a compactação do grupo ao qual $o$ pertence. Quanto menor o valor, mais compacto é o grupo. O valor de $b(o)$ reflete quanto o objeto $o$ está separado de outros grupos. Quanto maior o valor de $b(o)$, mais separado $o$ está dos outros grupos. Portanto, valores de coeficiente de silhueta próximos a 1 indicam que o grupo é compacto e está distante dos outros grupos, o que é desejável. Valores de coeficiente de silhueta negativos indicam que o objeto em questão está mais próximo de objetos de outros grupos do que dos objetos do grupo que ele pertence, o que não é desejável. Valores próximos de 0 podem indicar que o objeto em questão está entre dois grupos. Para avaliar a qualidade de um agrupamento, pode ser utilizado o valor médio do coeficiente de silhueta de todos os objetos no conjunto de dados (HAN; KAMBER; PEI, 2012).

### 5.4.2 Método de Kaplan-Meier e teste log-rank

Além da validação de agrupamento, o método de Kaplan-Meier e o teste log-rank são medidas utilizadas frequentemente para a análise da qualidade dos agrupamentos resultantes dos algoritmos. O Kaplan-Meier é um método simples e o mais popular utilizado para a análise de sobrevivência. Juntamente com o teste log-rank, é capaz de estimar as probabilidades de sobrevivência e comparar a sobrevivência entre os grupos (JAGER et al., 2008).

Em aplicações médicas, o método Kaplan-Meier é utilizado para medir a fração de indivíduos que vivem por um determinado período de tempo após o tratamento. O tempo a partir de um ponto definido até a ocorrência de um determinado evento é chamado de tempo de sobrevivência e a análise de dados de grupos é chamada de análise de sobrevivência. O evento analisado pode ser morte, ocorrência de uma infecção ou recidiva bioquímica de um câncer, por exemplo (BEWICK; CHEEK; BALL, 2004; GOEL; KHANNA; KISHORE, 2010). Entretanto, ao final do tempo de acompanhamento, é possível que o evento não tenha ocorrido com todos os pacientes do estudo. Para esses pacientes, o tempo de sobrevivência é considerado um dado censurado. O método Kaplan-Meier é capaz de calcular a sobrevivência ao longo do tempo mesmo com esse tipo de dado (JAGER et al., 2008).

Ao analisar os dados de sobrevivência, duas funções que dependem do tempo são calculadas: a função de sobrevivência e a função de risco. A função de sobrevivência $S(t)$ é definida como a probabilidade de o evento não ocorrer até o tempo $t$. A função de risco $h(t)$ é a probabilidade condicional de o evento ocorrer no tempo $t$ sendo que não ocorreu até aquele momento (BEWICK; CHEEK; BALL, 2004).

A curva de sobrevivência de Kaplan-Meier é definida como a probabilidade de sobrevivência $S(t)$ em um determinado período de tempo $t$. Considerando a morte por câncer como o evento de interesse, a probabilidade de sobrevivência em qualquer momento particular é calculada de acordo com a Equação 12.

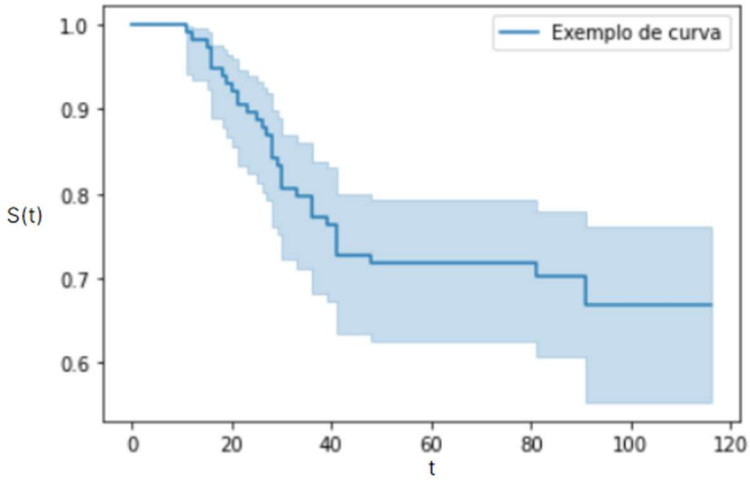
$$S_t = \frac{número\ de\ indivíduos\ vivos\ no\ início - número\ de\ óbitos}{número\ de\ indivíduos\ vivos\ no\ início} \qquad (12)$$

Para cada intervalo de tempo, a probabilidade de sobrevivência é calculada como o número de indivíduos sobreviventes dividido pelo número de indivíduos em risco. Indivíduos com dados censurados não são considerados indivíduos em risco. A probabilidade total de sobrevivência até esse intervalo de tempo é calculada multiplicando todas as probabilidades de

sobrevivência em todos os intervalos de tempo anteriores a esse tempo, aplicando a lei da multiplicação da probabilidade para calcular a probabilidade cumulativa (GOEL; KHANNA; KISHORE, 2010).

Duas ou mais curvas de sobrevivência podem ser comparadas visualmente, através do gráfico traçado entre as probabilidades de sobrevivência estimadas (eixo Y) e o tempo (eixo X) (Figura 6), ou estatisticamente testando a hipótese nula, ou seja, que não há diferença em relação à sobrevivência entre os grupos. Essa hipótese nula é testada estatisticamente por outro teste conhecido como teste log-rank (equação 13).

Figura 6- Exemplo de um gráfico de Kaplan-Meier com uma curva, mostrando a relação entre a probabilidade de sobrevivência estimada (S(t)) e o tempo (t). Como exemplo, considere que a curva representa um grupo de pacientes com câncer de próstata, que a probabilidade estimada seja de pacientes livres de recidiva bioquímica e que o tempo esteja em meses. Considerando 20 meses, aproximadamente 95% dos pacientes do grupo está livre de recidiva bioquímica. Já considerando 100 meses, temos aproximadamente 68% de pacientes livre de recidiva bioquímica.




Fonte: Elaborada pela autora.

$$log - rank = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} \qquad (13)$$

Considerando dois grupos distintos, $E_1$ e $E_2$ são o número esperado de eventos nos grupos 1 e 2, respectivamente, e $O_1$ e $O_2$ são o número total de eventos observados em cada grupo (BEWICK; CHEEK; BALL, 2004; GOEL; KHANNA; KISHORE, 2010).

Os cálculos de todos os valores na equação fornecerão o valor da estatística de teste. O p-valor pode ser calculado e analisado a partir dos valores obtidos. Quanto menor o p-valor, maior a diferenciação entre os grupos. Grupos com p-valor baixo (geralmente $< 0{,}05$ ) são

considerados grupos diferentes estatisticamente, ou seja, há diferença significativa entre os grupos quanto à sobrevida.

# 6 RESULTADOS E DISCUSSÃO

Com o objetivo de refinar os grupos de risco de câncer de próstata, as três estratégias descritas na seção 5.2 foram elaboradas, implementadas e analisadas com base no desempenho obtido, para identificação de pontos fortes e fracos e discussão sobre sua aplicabilidade. As três estratégias foram estudadas com a implementação de sete algoritmos de Aprendizado de Máquina não supervisionado e um algoritmo de Autoencoder, que utiliza técnica de aprendizado profundo não supervisionado. Os algoritmos são: K-means, Mini Batch K-means, Affinity Propagation, Agglomerative Clustering, BIRCH, DBSCAN, OPTICS e Autoencoder.

Inicialmente, antes da implementação dos algoritmos de agrupamento, os dados descritos na seção 5.1 foram analisados e pré-processados. Observou-se que, dos 485 pacientes, 81 pacientes tinham 2 ou mais informações faltantes na base de dados. Com isso, decidiu-se excluir estes pacientes, resultando em 404 pacientes no total para o estudo.

Em seguida, as variáveis que foram utilizadas para o agrupamento e para as análises posteriores foram selecionadas: idade, PSA inicial (PSAi), estágio clínico do tumor, escore de Gleason primário, escore de Gleason secundário, porcentagem positiva da biópsia (PPB), ocorrência de invasão perineural, presença de pressão alta, presença de diabetes, ocorrência da recidiva do PSA, o tempo em meses do seguimento do paciente e o grupo de risco no qual o paciente foi classificado previamente.

De todas as variáveis, idade, PSA, estágio clínico do tumor, escore de Gleason primário, escore de Gleason secundário, porcentagem positiva da biópsia, ocorrência de invasão perineural, se o paciente tem pressão alta e se o paciente tem diabetes, totalizando em 9 variáveis, foram selecionadas para serem utilizadas na implementação dos algoritmos para a realização dos agrupamentos. Das 9 variáveis, 3 são binárias (ocorrência de invasão perineural, se o paciente tem pressão alta e se o paciente tem diabetes) e as outras 6 foram normalizadas para que todos os valores estejam em uma mesma escala.

As variáveis restantes (ocorrência de recidiva do PSA, o tempo em meses do seguimento do paciente e o grupo de risco que o paciente foi classificado previamente) foram utilizadas depois do agrupamento para a implementação do gráfico de Kaplan-Meier e do teste log-rank. O gráfico mostra visualmente as curvas em relação ao tempo livre de recidiva bioquímica de cada grupo definido pelos algoritmos e o teste log-rank informa um p-valor a partir da comparação entre os grupos. Curvas bem separadas e um baixo p-valor indicam que os grupos são diferentes entre si. Tanto o gráfico quanto o p-valor resultantes foram comparados para que

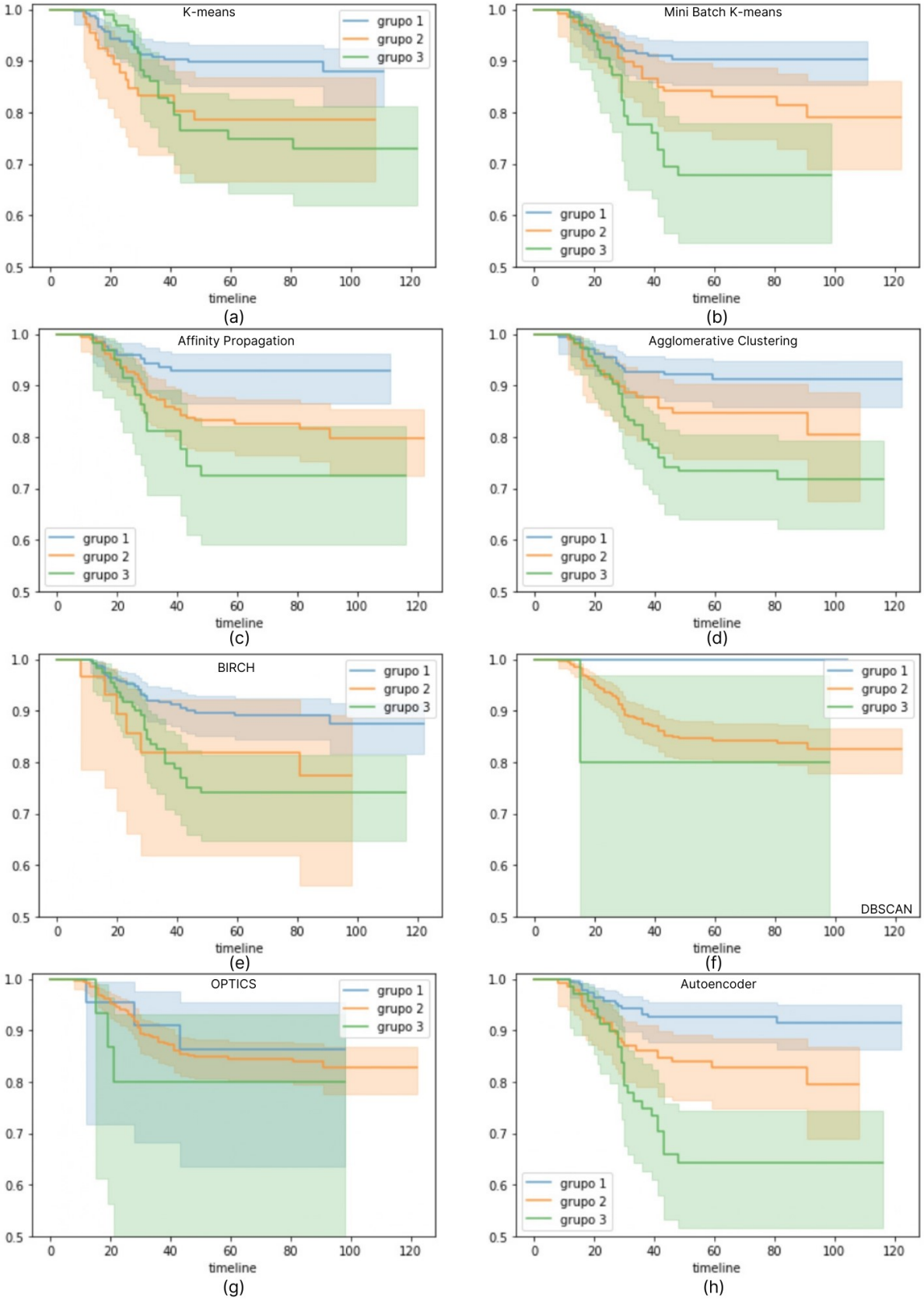
sejam obtidos os resultados mais satisfatórios entre os oito algoritmos implementados e entre as três estratégias elaboradas. Após as análises, para os melhores resultados obtidos, foi implementada árvore de decisão para que fossem identificados os indicadores mais relevantes para o agrupamento e os valores limites das variáveis para a classificação do paciente.

## 6.1 Estratégia 1

Após o pré-processamento dos dados, os algoritmos de aprendizado não supervisionado foram implementados para o agrupamento dos dados. A primeira implementação foi realizada de acordo com a Estratégia 1, que tem como objetivo formar três grupos de risco no total para refinar as condições de agrupamento. Os oito algoritmos foram implementados e os dados foram divididos de acordo com os grupos resultantes. Após o agrupamento, foram acrescentadas as variáveis ocorrência de recidiva do PSA e o tempo em meses do seguimento do paciente para a implementação do método de Kaplan-Meier, do teste log-rank e para o cálculo do valor do coeficiente de silhueta para a análise dos resultados.

Os gráficos de Kaplan-Meier resultantes dos oito algoritmos de Aprendizado de Máquina para o estudo da Estratégia 1 podem ser vistos na Figura 7. Os grupos 1, 2 e 3 representam, respectivamente, os grupos de risco baixo, intermediário e alto. Os valores resultantes do teste log-rank e do método da silhueta podem ser vistos na Tabela 12 e Tabela 13, respectivamente.

Figura 7 - Gráficos de Kaplan-Meier dos agrupamentos resultantes dos oito algoritmos, sendo (a) K-means, (b) Mini Batch K-means, (c) Affinity Propagation, (d) Agglomerative Clustering, (e) BIRCH, (f) DBSCAN, (g) OPTICS e (h) Autoencoder




Fonte: Elaborado pela autora.

Tabela 12 - p-valor obtido em cada algoritmo na comparação de pares do teste log-rank. É possível observar que os algoritmos que obtiveram os melhores resultados foram Mini Batch K-means, Agglomerative Clustering e Autoencoder.

| | $p - valor$ **(teste log-rank)** | | |
|---|---|---|---|
| **Algoritmo** | **Grupos 1 e 2** | **Grupos 1 e 3** | **Grupos 2 e 3** |
| K-means | **0,02** | **<0,005** | 0,69 |
| Mini Batch K-means | **0,02** | **<0,005** | **0,01** |
| Affinity Propagation | **0,01** | **<0,005** | 0,13 |
| Agglomerative Clustering | **0,05** | **<0,005** | **0,07** |
| BIRCH | 0,10 | **<0,005** | 0,74 |
| DBSCAN | 0,34 | 0,32 | 0,56 |
| OPTICS | 0,76 | 0,58 | 0,60 |
| Autoencoder | **0,01** | **<0,005** | **<0,005** |

Fonte: Elaborado pela autora.

Tabela 13 - Valores de coeficiente de silhueta obtidos em cada algoritmo, sendo os algoritmos DBSCAN e OPTICS os únicos com valores negativos.

| **Algoritmo** | **Coeficiente de Silhueta** |
|---|---|
| K-means | 0,18892 |
| Mini Batch K-means | 0,16497 |
| Affinity Propagation | 0,12314 |
| Agglomerative Clustering | 0,11519 |
| BIRCH | 0,18891 |
| DBSCAN | -0,36180 |
| OPTICS | -0,04130 |
| Autoencoder | 0,10743 |

Fonte: Elaborado pela autora.

Ao analisar os resultados dos gráficos de Kaplan-Meier e dos valores do método de silhueta, podemos observar que os algoritmos que utilizam métodos baseados em densidade apresentaram os resultados menos satisfatórios em comparação aos outros algoritmos. Tanto o DBSCAN quanto o OPTICS (Figura 7 (f) e (g)) não dividiram de forma que houvesse separação entre os três grupos e obtiveram um valor de coeficiente de silhueta negativo (-0, 36180 e -0, 04130, respectivamente), o que significa que os grupos não estão distantes um dos outros e não

são compactos. Isso pode ter ocorrido porque os algoritmos baseados em densidade não separaram de forma homogênea os pacientes nos três grupos. O algoritmo DBSCAN agrupou 394 pacientes em um único grupo, e 5 pacientes em cada grupo restante. Já o OPTICS agrupou 367 pacientes em um único grupo, 22 pacientes no segundo grupo e somente 15 pacientes no terceiro grupo.

O algoritmo de particionamento K-Means separou o grupo que corresponde ao de baixo risco dos outros dois grupos, os quais não foram muito bem definidos (Figura 7 (a)) e apresentou um valor de silhueta positivo (0,18892).

Os algoritmos Affinity Propagation e Agglomerative Clustering, apesar de utilizarem métodos diferentes de agrupamento, tiveram resultados semelhantes. Ambos separaram e definiram três grupos distintos (Figura 7(c) e 7(d)), mas com o p-valor indicando que os grupos intermediário e alto não foram tão bem separados quanto os grupos baixo e intermediário e baixo e alto. Ambos também apresentaram valores próximos de silhueta: 0,12314 e 0,11519, respectivamente.

Os resultados do algoritmo BIRCH mostram que os grupos de risco baixo e alto foram divididos de forma que ambos são diferentes e apresentam curvas distintas e separadas, porém pode ser observado nos gráficos de Kaplan-Meier que o grupo de risco intermediário não é muito bem definido, com sua curva posicionada entre as curvas dos grupos que representam os grupos baixo e alto risco (Figura 7 (e)). O coeficiente de silhueta também foi positivo (0,18891).

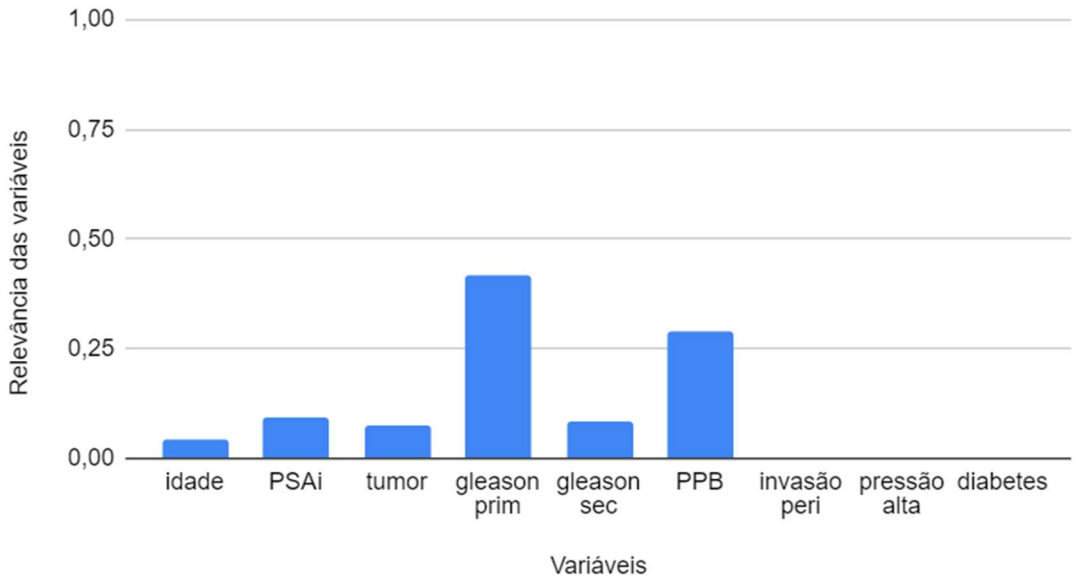
Por fim, os algoritmos Mini Batch K-means e Autoencoder apresentaram os menores valores de p-valor no teste log-rank nos três grupos de risco, o que indica maior separabilidade entre os grupos. É possível observar nos gráficos Kaplan-Meier resultantes destes dois algoritmos que os três grupos de riscos estão separados e apresentam tempos de sobrevivência distintos (Figura 7(b) e 7(h)). Além disso, os valores dos coeficientes de silhueta são positivos, ou seja, os grupos estão separados e compactos. Ambos os algoritmos geraram valores muito próximos, sendo os valores do teste de log-rank do Autoencoder levemente menores que os do Mini Batch K-means e o valor do coeficiente de silhueta do algoritmo Mini Batch K-Means maior do que o do algoritmo Autoencoder (0,16497 e 0,10743, respectivamente).

Diante destes resultados, o algoritmo escolhido como a melhor proposta nesta estratégia é o algoritmo de particionamento Mini Batch K-means. Os grupos baixo, intermediário e alto risco possuem 234, 110 e 60 pacientes, respectivamente. É possível observar que a divisão dos pacientes não foi feita de forma completamente homogênea, uma vez que apenas 60 pacientes foram classificados no grupo de alto risco e 234 foram classificados como baixo risco. Considerando o método D'Amico e seguindo os critérios da Tabela 2 apresentada na seção 5.1,

135 pacientes foram classificados no grupo de baixo risco, 148 no grupo intermediário e 121 no grupo de alto risco. Portanto, as duas classificações mostram divergências em relação ao número de pacientes em cada grupo de risco, sendo a classificação pelo método D'Amico mais homogênea em relação a classificação realizada pelo algoritmo Mini Batch K-means.

Após a escolha do algoritmo, foi realizada a análise da relevância de cada variável para o agrupamento realizado pelo Mini Batch K-means. Essa relevância foi obtida a partir da implementação da árvore de decisão, com o atributo feature_importances_, em que o algoritmo da árvore de decisão, após a implementação utilizando os dados resultantes do agrupamento, indica a relevância de cada variável. A Figura 8 ilustra que as variáveis de escore de Gleason primário e PPB foram as mais relevantes para o agrupamento realizado pelo algoritmo. Já as variáveis de porcentagem de pacientes com invasão perineural, pressão alta e diabetes não foram relevantes para o agrupamento, diferente das demais variáveis.

Figura 8 – Gráfico com a relevância das variáveis para o agrupamento realizado pelo algoritmo Mini Batch K-means. É possível observar o destaque da relevância das variáveis escore de Gleason primário e PPB.

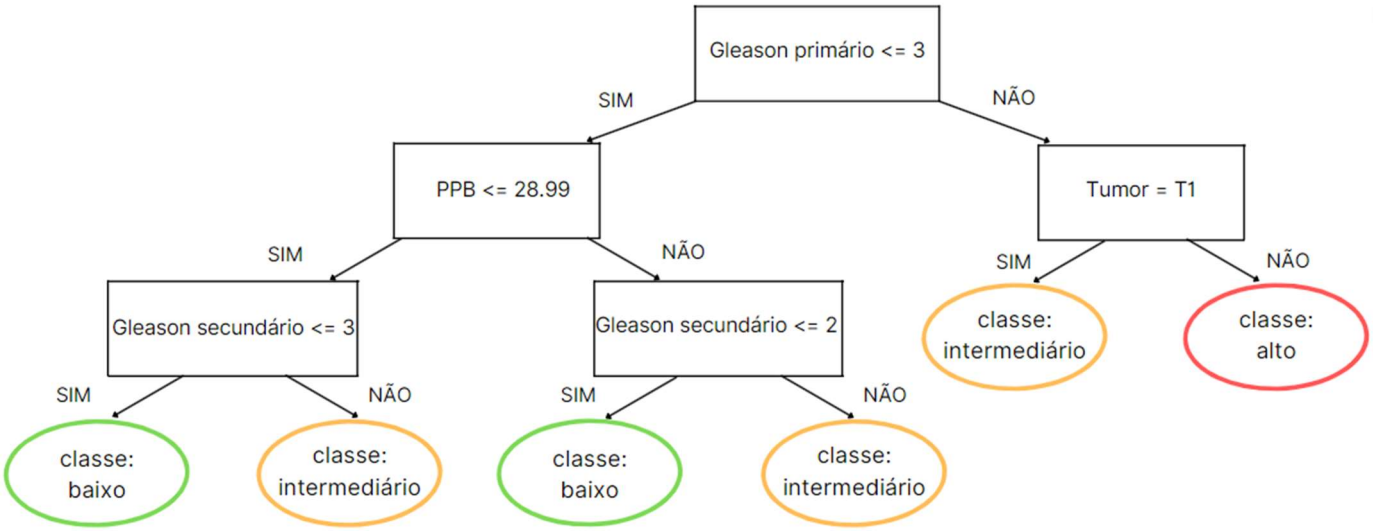



Fonte: Elaborado pela autora.

Por último, foi realizado a implementação da visualização de uma árvore de decisão para obter os valores de corte de cada variável para que um paciente seja classificado em um dos grupos de risco (Figura 9). Para a melhor visualização e compreensão do resultado, o limite definido de profundidade da árvore foi 3. Em cada nó é possível observar a variável em questão e o grupo de risco. A árvore de decisão indica que o primeiro valor a ser observado para o

agrupamento é o escore de Gleason primário. Se o valor for menor que 3, o paciente só poderá ser classificado nos grupos de risco baixo ou intermediário. Caso contrário, o paciente poderá ser classificado nos grupos de risco intermediário ou alto.

Figura 9 – Árvore de decisão implementada a partir do agrupamento realizado pelo algoritmo Mini Batch K-means, com profundidade igual a 3, informando a variável em questão, o valor da variável e o grupo de risco de cada nó.




Fonte: Elaborado pela autora.

Portanto, o algoritmo Mini Batch K-means é um algoritmo adequado para estabelecer três novos grupos de risco, refinando as regras de agrupamento de acordo com a árvore de decisão implementada. Todos os três grupos definidos pelo algoritmo são distintos e distantes entre si, pois apresentam valores baixos no teste de log-rank indicando que os grupos apresentam diferença significativa, valor de silhueta positivo, o que indica que os grupos são compactos e distantes entre si, e uma árvore de decisão coerente com os valores das variáveis utilizadas para o agrupamento, sendo os menores valores pertencentes ao grupo de baixo risco e os maiores valores pertencentes ao grupo de alto risco.

Para classificar o paciente em um dos três grupos de risco, diferente dos critérios utilizados pelo método D'Amico que considera escore de Gleason total, estágio clínico do tumor e valor de PSAi, a classificação resultante do algoritmo Mini Batch K-means considera as variáveis escore de Gleason primário, PPB, escore de Gleason secundário e estágio clínico do tumor. Sendo que o escore de Gleason total é a soma das variáveis escore de Gleason primário e secundário, há divergência entre os métodos apenas em relação às variáveis PSAi e PPB. Já em relação aos valores das variáveis, é possível observar uma semelhança entre os

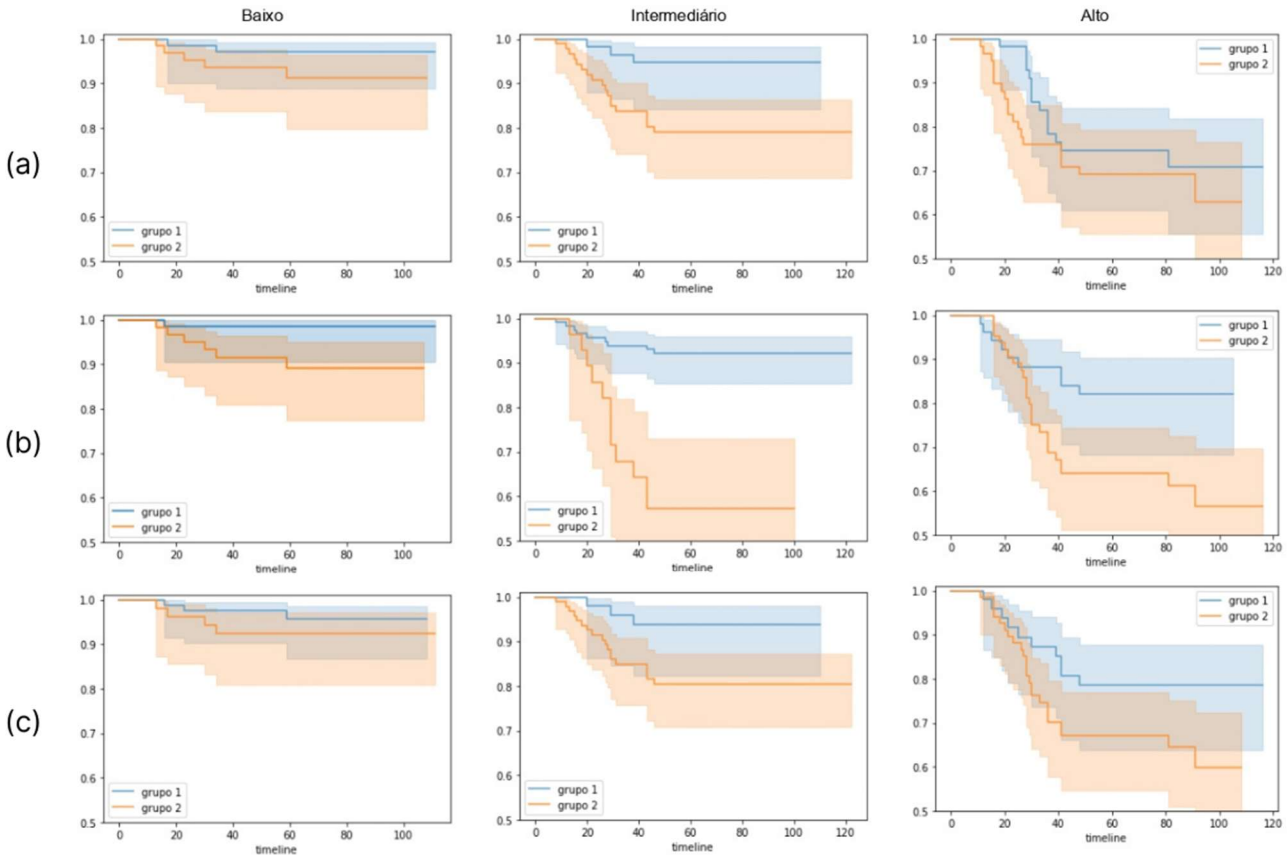
métodos em relação ao escore de Gleason: tanto na classificação D'Amico quanto na classificação resultante do algoritmo de agrupamento, só é possível ser classificado como baixo risco se o escore de Gleason total for menor que 7.

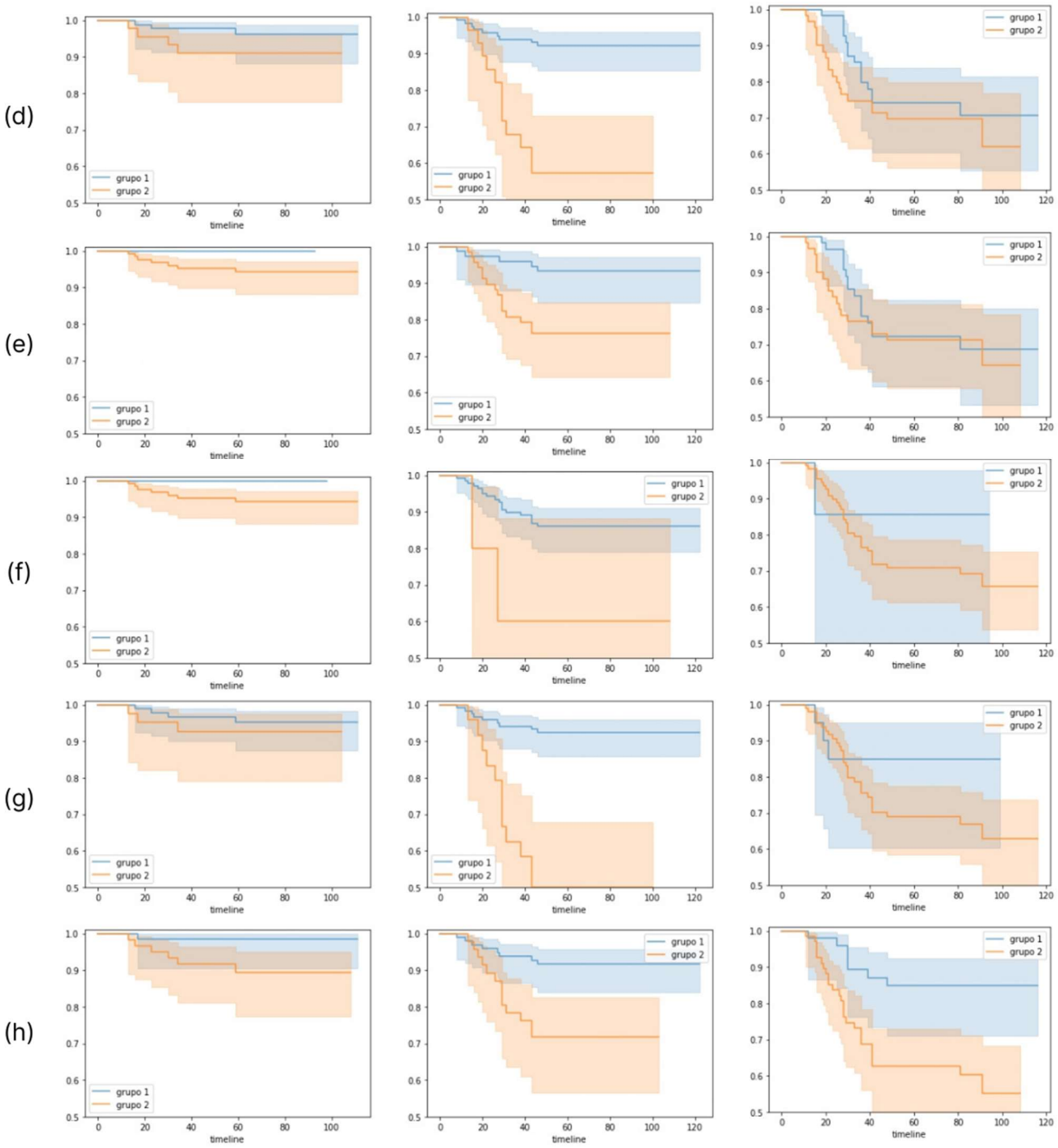
## 6.2 Estratégia 2

A segunda estratégia elaborada para o presente estudo com o objetivo de refinar os grupos de risco do câncer de próstata consiste em formar subgrupos a partir dos grupos de risco já definidos pelo método D'Amico. Os 404 pacientes foram previamente classificados em grupos de baixo, intermediário e alto risco, sendo 135 pacientes no grupo de baixo risco, 148 no grupo intermediário e 121 no grupo de alto risco, de acordo com os critérios da Tabela 1.

Novamente os oito algoritmos foram utilizados, agora com o número de grupos definidos igual a 2. Cada grupo de risco foi processado de forma separada, para que os subgrupos fossem definidos. Os gráficos de Kaplan-Meier resultantes podem ser observados na Figura 10 e os resultados de p-valor do teste log-rank e do coeficiente de silhueta podem ser observados nas Tabelas 14 e 15, respectivamente.

Figura 10 - Gráficos de Kaplan-Meier dos agrupamentos resultantes para formar os subgrupos a partir dos grupos de risco baixo, intermediário e alto, respectivamente, sendo os algoritmos (a) K-means, (b) Mini Batch K-means, (c) Affinity Propagation, (d) Agglomerative Clustering, (e) BIRCH, (f) DBSCAN, (g) OPTICS e (h) Autoencoder






Fonte: Elaborado pela autora.

Tabela 14 - p-valor obtido em cada algoritmo no teste log-rank. Os melhores resultados foram obtidos pelos algoritmos Mini Batch K-means e Autoencoder.

| | $p - valor$ (teste log-rank) | | |
|---|---|---|---|
| **Algoritmo** | **Baixo** | **Intermediário** | **Alto** |
| K-means | 0,18 | **0,01** | 0,40 |
| Mini Batch K-Means | **0,03** | **<0,005** | **0,02** |

| | | | |
|---|---|---|---|
| Affinity Propagation | 0,35 | **0,03** | 0,11 |
| Agglomerative Clustering | 0,18 | **<0,005** | 0,42 |
| BIRCH | 0,59 | **<0,005** | 0,72 |
| DBSCAN | 0,60 | **0,06** | 0,42 |
| OPTICS | 0,50 | **<0,005** | 0,15 |
| Autoencoder | **0,02** | **<0,005** | **<0,005** |

Fonte: Elaborado pela autora.

Tabela 15 - Valores de coeficiente de silhueta obtidos em cada algoritmo, sendo todos os valores positivos, exceto pelos valores obtidos pelo algoritmo DBSCAN.

| | **Coeficiente de Silhueta** | | |
|---|---|---|---|
| **Algoritmo** | **Baixo** | **Intermediário** | **Alto** |
| K-means | 0,17776 | 0,14496 | 0,21902 |
| Mini Batch K-Means | 0,16156 | 0,17741 | 0,21778 |
| Affinity Propagation | 0,19513 | 0,13846 | 0,19747 |
| Agglomerative Clustering | 0,18340 | 0,17741 | 0,20345 |
| BIRCH | 0,48729 | 0,14150 | 0,21202 |
| DBSCAN | -0,14179 | -0,19709 | -0,02411 |
| OPTICS | 0,16419 | 0,15077 | 0,09338 |
| Autoencoder | 0,17230 | 0,17741 | 0,11300 |

Fonte: Elaborado pela autora.

Observando o desempenho comparativo entre os algoritmos, o DBSCAN, novamente, teve o resultado menos satisfatório. Foi o único algoritmo a obter resultados negativos no método de silhueta (-0, 14179, -0,19709 e -0,02411 nos subgrupos definidos a partir dos grupos de baixo, intermediário e alto risco, respectivamente). Assim como na Estratégia 1, o algoritmo não dividiu de forma homogênea os subgrupos. Dos 135 pacientes pertencentes ao grupo de baixo risco, 5 foram agrupados no subgrupo 1 e 130 no subgrupo 2. Os 148 pacientes do grupo de risco intermediário foram subdivididos de forma que 5 foram agrupados no subgrupo 1 e 143 no subgrupo 2. Por último, os 121 pacientes pertencentes ao grupo de risco alto foram subdivididos de forma que 7 pacientes foram agrupados no subgrupo 1 e 114 no subgrupo 2. Outra observação feita foi que o algoritmo não classificou pacientes que apresentaram recidiva bioquímica nos subgrupos 1 do grupo de risco baixo, o que é representado visualmente no

gráfico de Kaplan-Meier através da linha reta na Figura 7(f). O mesmo acontece no agrupamento do algoritmo BIRCH, também no grupo de baixo risco (Figura 7(e)).

No geral, com exceção dos algoritmos Mini Batch K-means e Autoencoder, os algoritmos não tiveram um bom desempenho para subdividir os pacientes classificados nos grupos de baixo e alto risco. Apesar dos valores de coeficiente de silhueta obtidos serem positivos, os valores obtidos a partir do teste log-rank foram altos, o que indica que não há diferença significativa entre os subgrupos definidos. Já em relação aos subgrupos criados a partir do grupo intermediário, é possível observar que todos os algoritmos definiram grupos diferentes entre si, com curvas bem definidas e separadas no gráfico de Kaplan-Meier e com baixos valores no teste log-rank. Como o grupo de risco intermediário é considerado o grupo mais heterogêneo, é muito importante ter como resultado subgrupos a partir deste grupo de risco.

### 6.2.1 Baixo risco

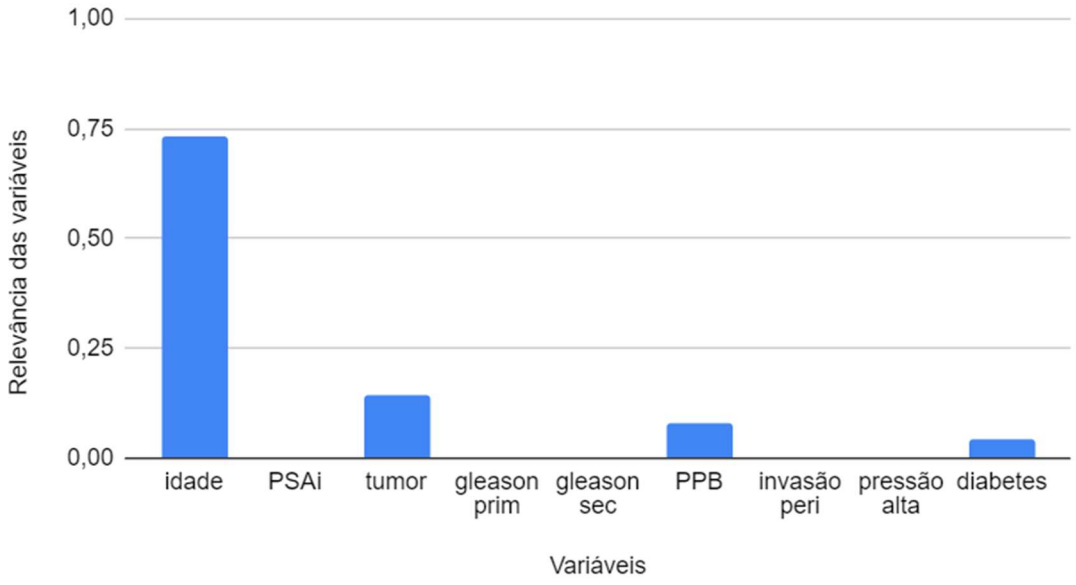
Para a classificação no grupo de baixo risco, os critérios do método D'Amico que devem ser cumpridos são: escore de Gleason total menor que 7, estágio clínico do tumor T1 ou T2a e valor de PSAi menor que 10 ng/mL. Como a presente estratégia tem como objetivo subdividir os grupos de risco já existentes pelo método D'Amico, todos os dados dos pacientes utilizados nesta abordagem de subdividir o grupo de baixo risco correspondem a estes critérios.

Como resultado, tanto o algoritmo Mini Batch K-means quanto o Autoencoder obtiveram grupos separados no gráfico de Kaplan-Meier (Figura 7(b) e 7(h)), p-valor baixo (0,03 e 0,02, respectivamente) e coeficiente de silhueta positivo (0,16156 e 0,17230). Diante disso, o algoritmo escolhido como o mais apropriado para realizar essa subdivisão é o Autoencoder.

O algoritmo classificou 76 dos 135 pacientes no subgrupo 1, que representa o grupo de risco muito baixo, e 59 no subgrupo 2, que representa o subgrupo de risco baixo, mais próximo do grupo de risco intermediário.

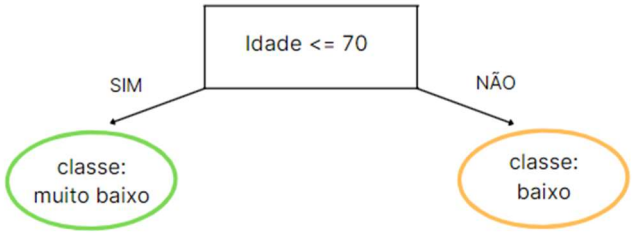
Entre as variáveis, é possível observar na Figura 11 que a variável idade é a mais relevante para o agrupamento e que as variáveis PSA, escore de Gleason primário e secundário, invasão perineural e pressão alta são as menos relevantes. Ao implementar a árvore de decisão, a relevância da variável idade fica em evidência, uma vez que a árvore é composta apenas por essa variável (Figura 12). O valor de corte da variável para definir se um novo paciente a ser classificado pertence ao grupo de risco muito baixo ou baixo é 70 anos.

Figura 11 – Gráfico com a relevância das variáveis para a divisão do grupo de baixo risco realizado pelo algoritmo Autoencoder. A relevância da variável idade se destaca em relação as outras variáveis.




Fonte: Elaborado pela autora.

Figura 12 – Árvore de decisão implementada a partir do agrupamento realizado pelo algoritmo Autoencoder utilizando dados dos pacientes do grupo de risco baixo, informando a variável em questão, o valor da variável e o grupo de risco de cada nó.




Fonte: Elaborado pela autora.

Diante disto, o algoritmo Autoencoder apresenta resultados muito satisfatórios para definir dois subgrupos de risco a partir do grupo de risco baixo. Todos os valores obtidos indicam que os subgrupos são distintos e distantes entre si, além de obter uma árvore de decisão apenas com a variável idade para que um paciente seja classificado em um dos subgrupos.

Obter uma árvore de decisão com apenas uma variável é importante para que o método seja facilmente utilizado, uma vez que esta classificação é realizada após a classificação de acordo com os critérios de D'Amico. Portanto, após o paciente ser classificado no grupo de baixo risco seguindo as regras do método de D'Amico, o mesmo pode ser classificado em um subgrupo considerando apenas sua idade.

### 6.2.2 Risco intermediário

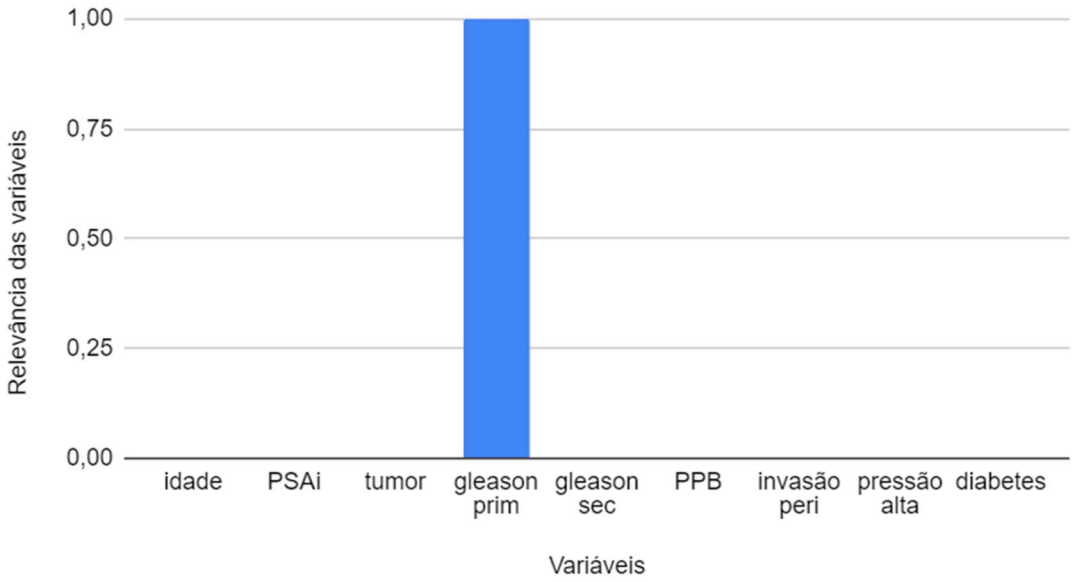
Para a classificação do grupo de risco intermediário, de acordo com o método D'Amico, o escore de Gleason deve ser igual a 7 ou o estágio clínico do tumor ser igual a T2b ou o valor de PSAi estar entre 10 e 20 ng/mL. Como a presente estratégia tem como objetivo subdividir os grupos de risco já existentes pelo método D'Amico, todos os dados dos pacientes utilizados nesta abordagem de subdividir o grupo de risco intermediário correspondem a estes critérios.

O grupo de risco intermediário, quando processado por todos os algoritmos, apresentou dois subgrupos bem definidos. Os melhores valores do teste log-rank foram obtidos pelos algoritmos Mini Batch K-means, Agglomerative Clustering, BIRCH, OPTICS e Autoencoder (todos com p-valor < 0,005). Já os melhores resultados do coeficiente de silhueta foram obtidos pelos algoritmos Mini Batch K-means, Agglomerative Clustering e Autoencoder (todos com o valor de 0,17741). Como o algoritmo de Autoencoder foi considerado o mais apropriado para realizar a subdivisão do grupo de baixo risco, foi decido considerá-lo novamente para o grupo de risco intermediário.

Dos 148 pacientes classificados como risco intermediário, o algoritmo classificou 119 no subgrupo 1, que representa o subgrupo de risco intermediário baixo, o mais próximo do grupo de risco baixo, e 29 no subgrupo 2, que representa o subgrupo de risco intermediário alto, o mais próximo do grupo de risco alto.

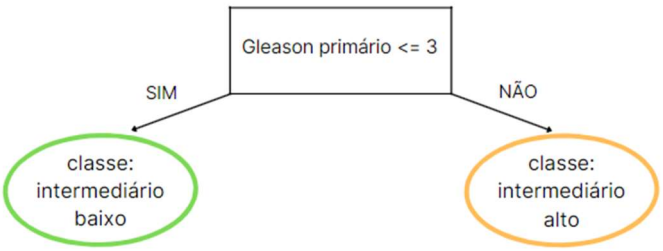
O gráfico de relevância das variáveis obtida a partir da implementação da árvore de decisão (Figura 13) ilustra que a variável mais relevante para o agrupamento é o escore de Gleason primário e não considera nenhuma outra variável, assim como a árvore de decisão (Figura 14), que mostra o valor de corte 3 para classificar um paciente no subgrupo de risco intermediário baixo ou intermediário alto. Após ser classificado no grupo de risco intermediário segundo as regras do método D'Amico, se o valor de escore de Gleason primário for menor ou igual a 3, o paciente é classificado como risco intermediário baixo; caso contrário, é classificado como risco intermediário alto.

Figura 13 – Gráfico com a relevância das variáveis para a divisão do grupo de risco intermediário realizado pelo algoritmo Autoencoder. Apenas a variável escore de Gleason primário foi considerada relevante para o agrupamento.




Fonte: Elaborado pela autora.

Figura 14 – Árvore de decisão implementada a partir do agrupamento realizado pelo algoritmo Autoencoder utilizando dados dos pacientes do grupo de risco intermediário. Se o escore de Gleason primário for menor ou igual a 3, o paciente é classificado no subgrupo intermediário baixo; caso contrário, é classificado no subgrupo intermediário alto.



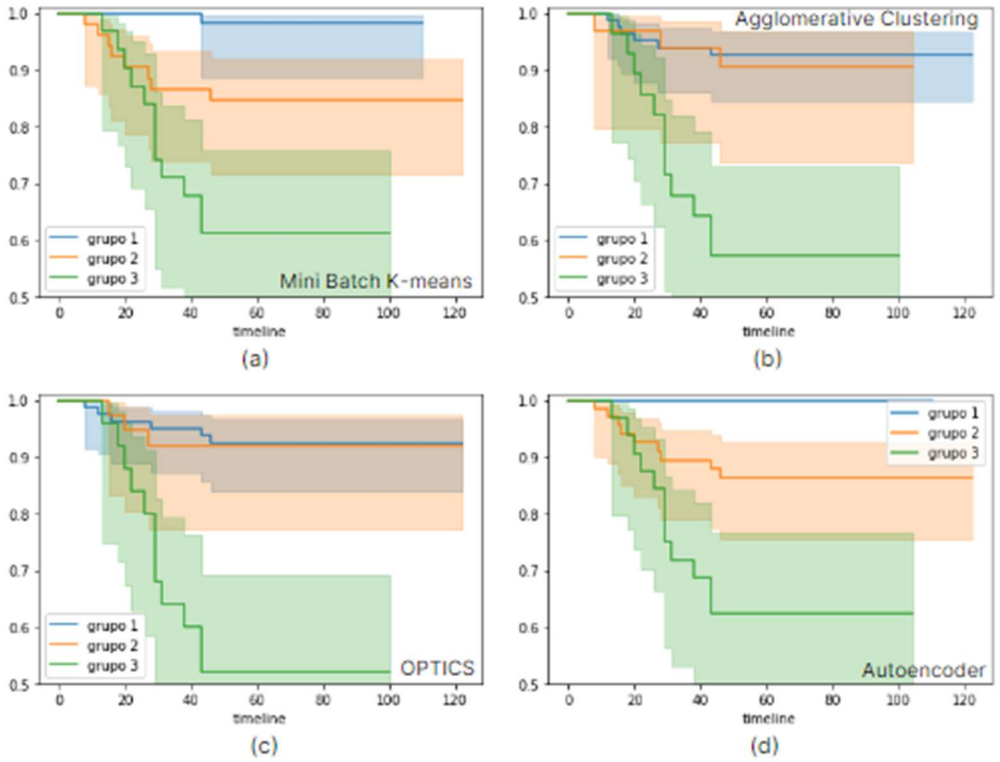

Fonte: Elaborado pela autora.

Com isso, o Autoencoder, novamente, é um algoritmo adequado para estabelecer dois subgrupos de risco a partir do grupo de risco intermediário. Os subgrupos intermediário baixo e intermediário alto formados são distantes entre si de acordo com o gráfico de Kaplan-Meier, diferentes de forma significativa de acordo com os valores do teste de log-rank e compactos de acordo com o coeficiente de silhueta. Além disso, o agrupamento também apresenta uma árvore de decisão coerente e que pode ser facilmente utilizada, com apenas uma variável definindo o valor limite para classificar um paciente em um dos subgrupos.

#### 6.2.2.1 Risco intermediário com 3 subgrupos

É possível observar nos gráficos de Kaplan Meier dos algoritmos com os melhores resultados no teste de log-rank (Mini Batch K-means, Agglomerative Clustering, BIRCH, OPTICS e Autoencoder) que ao dividir o grupo de risco intermediário em dois subgrupos, eles podem ficar muito distantes um do outro. Assim, foi realizada também a subdivisão do grupo de risco intermediário em três subgrupos utilizando os algoritmos que apresentaram os gráficos com maior distância entre os dois subgrupos: Mini Batch K-means, Agglomerative Clustering, OPTICS e Autoencoder. Os gráficos de Kaplan-Meier resultantes podem ser vistos na Figura 15 e os valores do teste log-rank e do coeficiente de silhueta podem ser vistos nas Tabelas 16 e 17, respectivamente.

Figura 15 – Gráficos de Kaplan-Meier dos agrupamentos resultantes para formar três subgrupos a partir dos grupos de risco intermediário, respectivamente, sendo os algoritmos (a) Mini Batch K-means, (b) Agglomerative Clustering, (c) OPTICS e (d) Autoencoder.




Fonte: Elaborado pela autora.

Tabela 16 - p-valor obtido em cada algoritmo na comparação de pares do teste log-rank. Os melhores resultados foram obtidos pelos algoritmos Mini Batch K-means e Autoencoder.

| | $p - valor$ (teste log-rank) | | |
|---|---|---|---|
| **Algoritmo** | **Grupos 1 e 2** | **Grupos 1 e 3** | **Grupos 2 e 3** |
| Mini Batch K-Means | **0,01** | **<0,005** | **0,03** |
| Agglomerative Clustering | 0,72 | **<0,005** | **<0,005** |
| OPTICS | 0,94 | **<0,005** | **<0,005** |
| Autoencoder | **0,01** | **<0,005** | **0,01** |

Fonte: Elaborado pela autora.

Tabela 17 - Valores de coeficiente de silhueta obtidos em cada algoritmo, todos positivos.

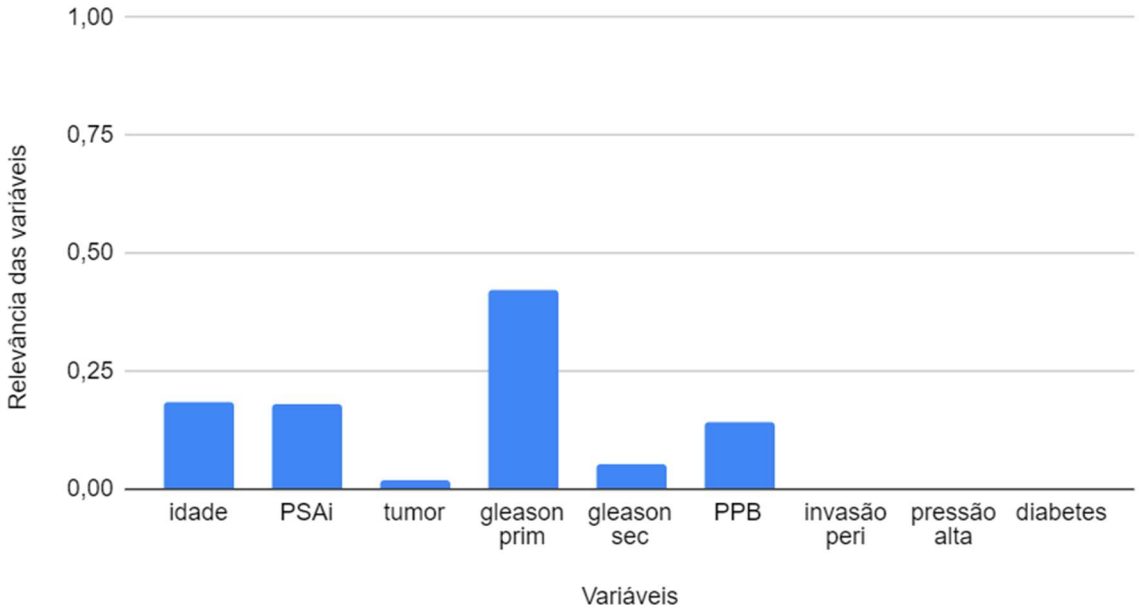
| **Algoritmo** | **Coeficiente de Silhueta** |
|---|---|
| Mini Batch K-Means | 0,15810 |
| Agglomerative Clustering | 0,19133 |
| OPTICS | 0,10421 |
| Autoencoder | 0,14474 |

Fonte: Elaborado pela autora.

Os gráficos de Kaplan-Meier e os valores obtidos pelo teste log-rank mostram que os algoritmos Mini Batch K-means e Autoencoder obtiveram os resultados mais satisfatórios ao dividir o grupo de risco intermediário em três subgrupos. Os números de p-valor do Autoencoder são levemente menores que os valores do Mini Batch K-means e os valores do coeficiente de silhueta são muito semelhantes. Já os algoritmos Agglomerative Clustering e OPTICS definiram muito bem um subgrupo de risco, que representa o intermediário alto, mas não definiram os outros dois subgrupos de forma distinta e distantes entre si. Diante disto, continuaremos com a análise considerando o algoritmo Autoencoder o mais adequado para esta divisão.

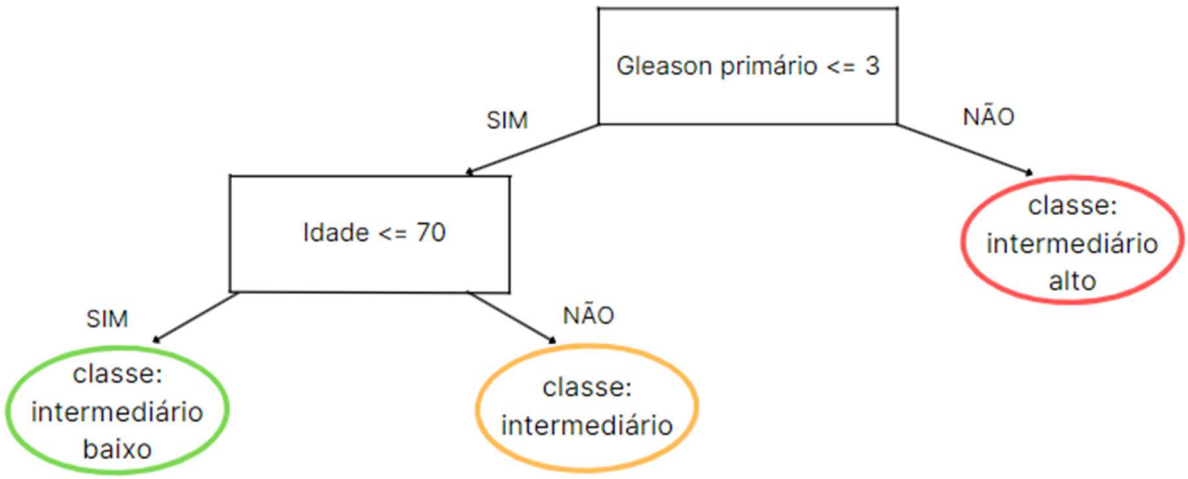
A Figura 16 mostra que, para este agrupamento, as variáveis mais relevantes são o escore de Gleason primário, seguido por idade, PSAi e PPB. Já a Figura 17 mostra a árvore de decisão implementada a partir do agrupamento. A árvore evidencia a relevância das variáveis escore de Gleason primário e idade, já que observando apenas os valores destas duas variáveis já é possível realizar a classificação. Na árvore, a classe “baixo” refere-se ao subgrupo mais próximo do grupo de risco baixo (S1), a classe “intermediário” o subgrupo S2 e a classe “alto” refere-se ao subgrupo mais próximo do grupo de risco alto (S3).

Figura 16 – Gráfico com a relevância das variáveis para a divisão do grupo de risco intermediário em três subgrupos realizado pelo algoritmo Autoencoder. A relevância da variável escore de Gleason primário se destaca em relação as outras variáveis.




Fonte: Elaborado pela autora.

Figura 17 – Árvore de decisão implementada a partir dos três subgrupos do grupo de risco intermediário formado pelo algoritmo Autoencoder.




Fonte: Elaborado pela autora.

Diante destes resultados, é possível observar que outra opção viável para a subdivisão do grupo de risco intermediário é utilizar o algoritmo de Autoencoder para dividir o grupo em questão em três subgrupos: intermediário baixo (o subgrupo mais próximo do grupo de risco baixo), intermediário e intermediário alto (o subgrupo mais próximo do grupo de risco alto).

Esses resultados são importantes, uma vez que o grupo de risco intermediário é considerado o mais heterogêneo entre os grupos.

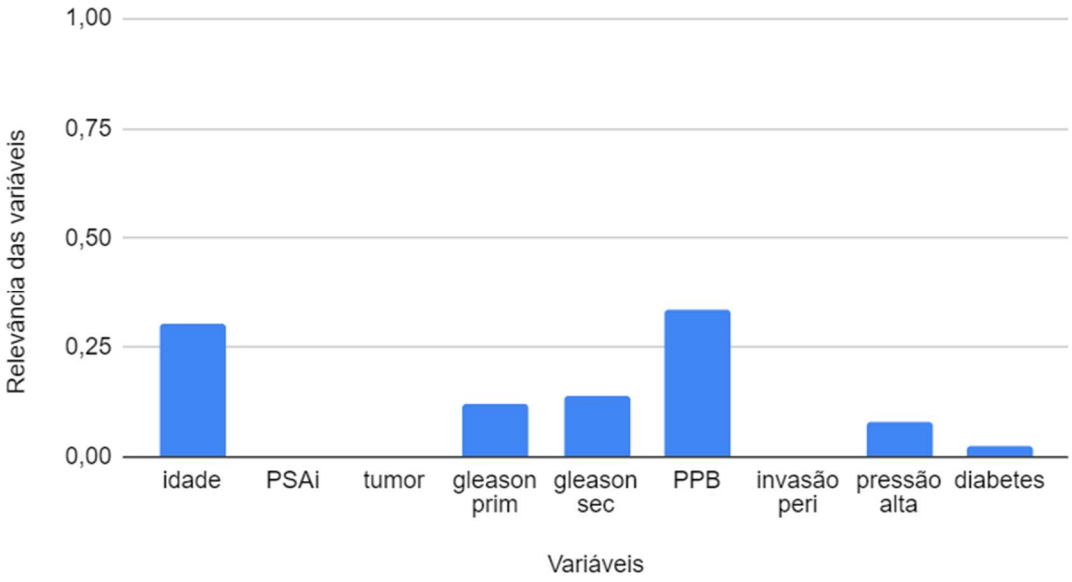
### 6.2.3 Alto risco

Por último, para classificação no grupo de alto risco seguindo as regras de D'Amico, o valor do escore de Gleason total deve ser maior que 7, ou o estágio clínico do tumor maior que T2b ou o valor de PSAi maior que 20ng/mL. Como a presente estratégia tem como objetivo subdividir os grupos de risco já existentes pelo método D'Amico, todos os dados dos pacientes utilizados nesta abordagem de subdividir o grupo de risco intermediário correspondem a estes critérios.

Assim como na divisão em dois subgrupos do grupo de risco intermediário, para definir dois subgrupos a partir do grupo de alto risco, os algoritmos que apresentaram os resultados mais satisfatórios foram Mini Batch K-means e Autoencoder. Os valores obtidos pelos algoritmos Mini Batch K-means e Autoencoder no teste log-rank foram 0,02 e <0,005, respectivamente. Os valores de coeficiente de silhueta de ambos foram positivos (0,21778 e 0,11300, respectivamente). Diante disto, o algoritmo Autoencoder também foi considerado o mais apropriado para realizar a subdivisão deste grupo.

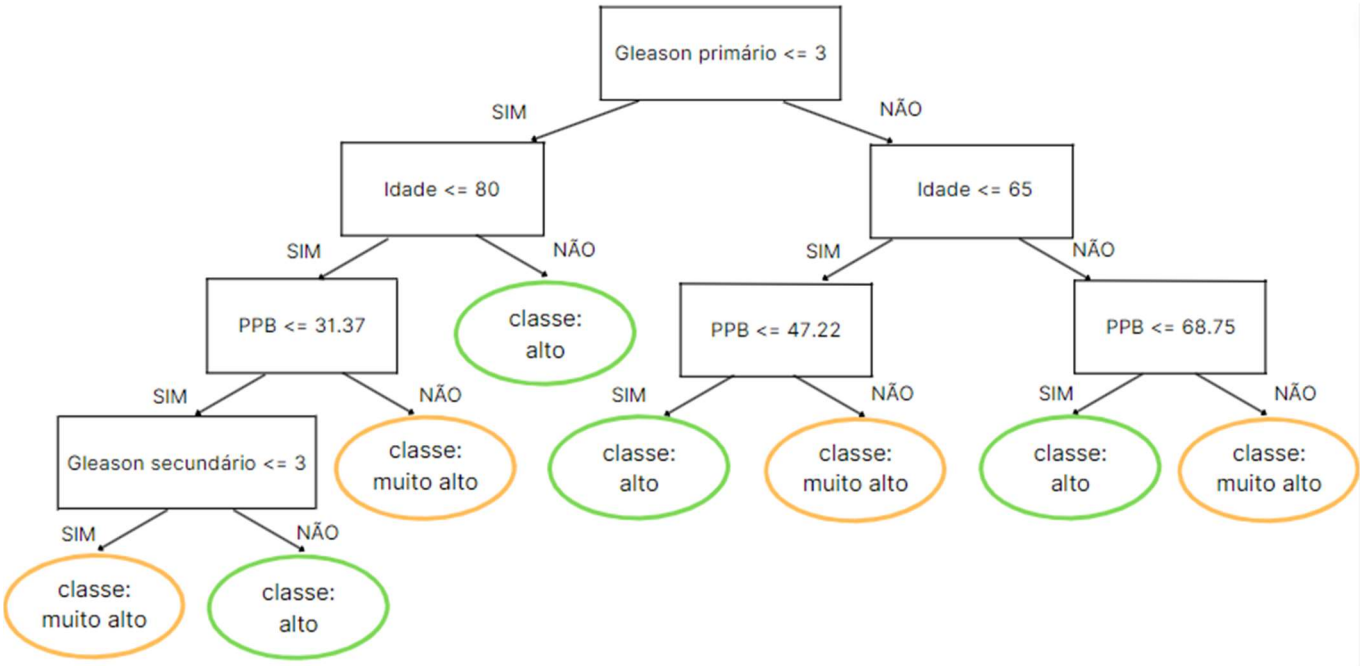
O algoritmo classificou 49 pacientes no subgrupo 1, subgrupo de risco alto, mais próximo do grupo intermediário, e 72 pacientes no subgrupo 2, subgrupo de risco muito alto. Entretanto, é possível observar no gráfico que ilustra a relevância de cada variável para o agrupamento (Figura 18) que as variáveis em destaque são PPB e idade. Já a árvore de decisão implementada (Figura 19), considera o escore de Gleason primário como a primeira variável a ser considerada na classificação de um paciente. Essas inconsistências em relação ao papel de cada variável na realização da classificação dos pacientes nos subgrupos criados a partir do grupo de risco alto indica um viés nos dados utilizados. Com isso, não foi obtido um resultado apropriado para a classificação dos pacientes de alto risco em dois subgrupos distintos.

Figura 18 – Gráfico com a relevância das variáveis para a divisão do grupo de alto risco realizado pelo algoritmo Autoencoder. As variáveis PPB e idade são as variáveis com mais relevância para o agrupamento.




Fonte: Elaborado pela autora.

Figura 19 – Árvore de decisão construída a partir dos dois subgrupos do grupo de risco alto formado pelo algoritmo Autoencoder.

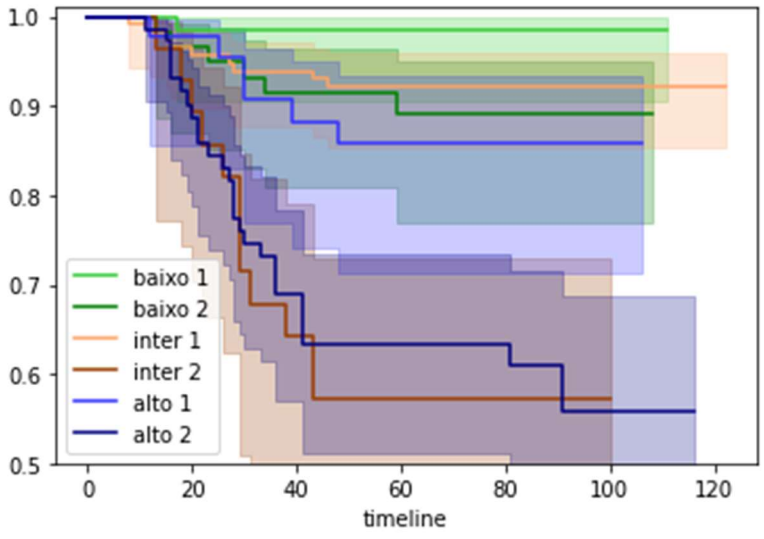



Fonte: Elaborado pela autora.

### 6.2.4 Acoplamento dos subgrupos

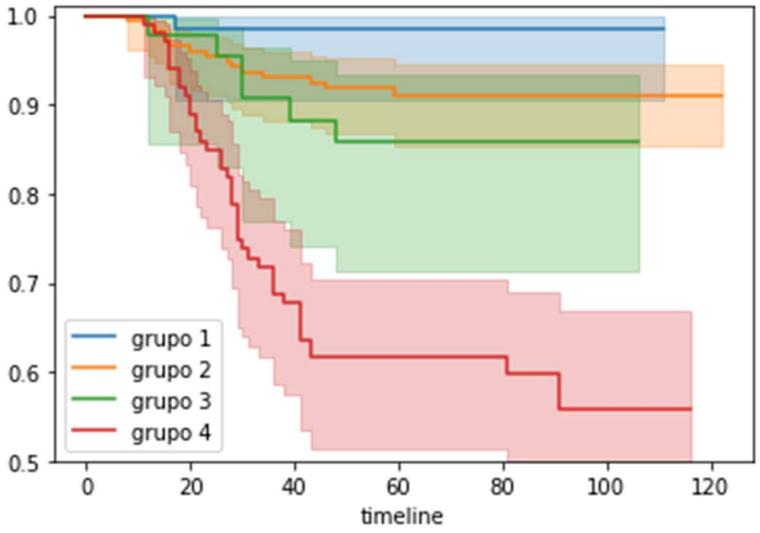
Para análises complementares, foi implementado também um gráfico de Kaplan-Meier e aplicado o teste log-rank e coeficiente de silhueta com todos os subgrupos criados na Estratégia 2. Ou seja, os subgrupos criados pelo algoritmo Autoencoder, que foi definido como o algoritmo mais apropriado para realizar o agrupamento dos dados em todos os grupos de risco desta estratégia, a partir dos grupos de risco baixo, intermediário e alto, foram representados em apenas um gráfico para analisarmos o comportamento de cada subgrupo. Este resultado pode ser visto na Figura 20. É possível observar que subgrupos criados a partir de grupos de risco diferentes se sobrepõem quando colocados em um mesmo gráfico. As curvas sobrepostas foram acopladas em apenas um grupo e um novo gráfico foi criado (Figura 21). Assim, os subgrupos com características semelhantes e que tinham curvas sobrepostas no gráfico Kaplan-Meier formaram apenas um grupo de risco. Isso ocorreu com os subgrupos baixo 2 (subgrupo do grupo de baixo risco mais próximo do grupo de risco intermediário) e inter 1 (subgrupo do grupo de risco intermediário mais próximo do grupo de risco baixo) e também com inter 2 (subgrupo do grupo de risco intermediário mais próximo do grupo de risco alto) e alto 2 (subgrupo do grupo de risco alto que representa o risco muito alto). Com isso, foram obtidos 4 grupos de risco no total. Os resultados do teste log-rank e método de silhueta para esses grupos podem ser vistos na Tabela 18.

Figura 20 – Gráfico de Kaplan-Meier com as curvas de todos os subgrupos obtidos pelo algoritmo Aprendizado Profundo a partir dos três grupos de riscos já definidos. É possível observar que as curvas dos subgrupos baixo 2 (verde escuro) e inter 1 (laranja) se sobrepõem, assim como as curvas dos subgrupos inter 2 (marrom) e alto 2 (azul escuro).




Fonte: Elaborado pela autora.

Figura 21 – Gráfico de Kaplan-Meier com quatro grupos, após o acoplamento dos subgrupos que apresentavam curvas sobrepostas.




Fonte: Elaborado pela autora.

Tabela 18 - p-valor e coeficiente de silhueta obtidos a partir dos quatro grupos de risco formados pelos subgrupos definidos pelo algoritmo Autoencoder.

| *p-valor* (teste log-rank) | | | | | | Coeficiente de silhueta |
|---|---|---|---|---|---|---|
| **Grupos 1 e 2** | **Grupos 1 e 3** | **Grupos 1 e 4** | **Grupos 2 e 3** | **Grupos 2 e 4** | **Grupos 3 e 4** | |
| **0,04** | **0,01** | **<0,005** | 0,32 | **<0,005** | **<0,005** | -0,04756 |

Fonte: Elaborado pela autora.

É possível observar que apesar de os grupos estarem visivelmente separados no gráfico de Kaplan-Meier, o valor de coeficiente de silhueta é negativo, indicando que os grupos não estão compactos e distantes um dos outros. Portanto, a abordagem de juntar os subgrupos para formar quatro grupos de risco não é a abordagem mais adequada. Além disso, o único grupo de risco que teve seus dois subgrupos acoplados com outros subgrupos foi o grupo de risco intermediário. Isso indica que o comportamento dos dados pertencentes ao grupo de risco intermediário é similar aos outros subgrupos de baixo e alto risco. O resultado de outra abordagem para obter quatro grupos de risco pode ser visto na seção a seguir.

## 6.3 Estratégia 3

A última estratégia foi implementada com o objetivo de definir no mínimo 4 grupos de risco com os dados dos pacientes disponíveis. A cada novo processamento de cada algoritmo

de agrupamento, o número de grupos a serem estabelecidos foi aumentado, iniciando por 4 grupos, até que não foram mais identificados novos grupos no gráfico de Kaplan-Meier, ou seja, até que não foram mais identificadas diferenças no tempo livre de recidiva bioquímica entre os novos grupos. Essa diferença entre os grupos não foi identificada a partir de 5 grupos. Portanto, os resultados apresentados serão dois: um agrupamento definindo 4 grupos de risco e um agrupamento definindo 5 grupos de risco no total.

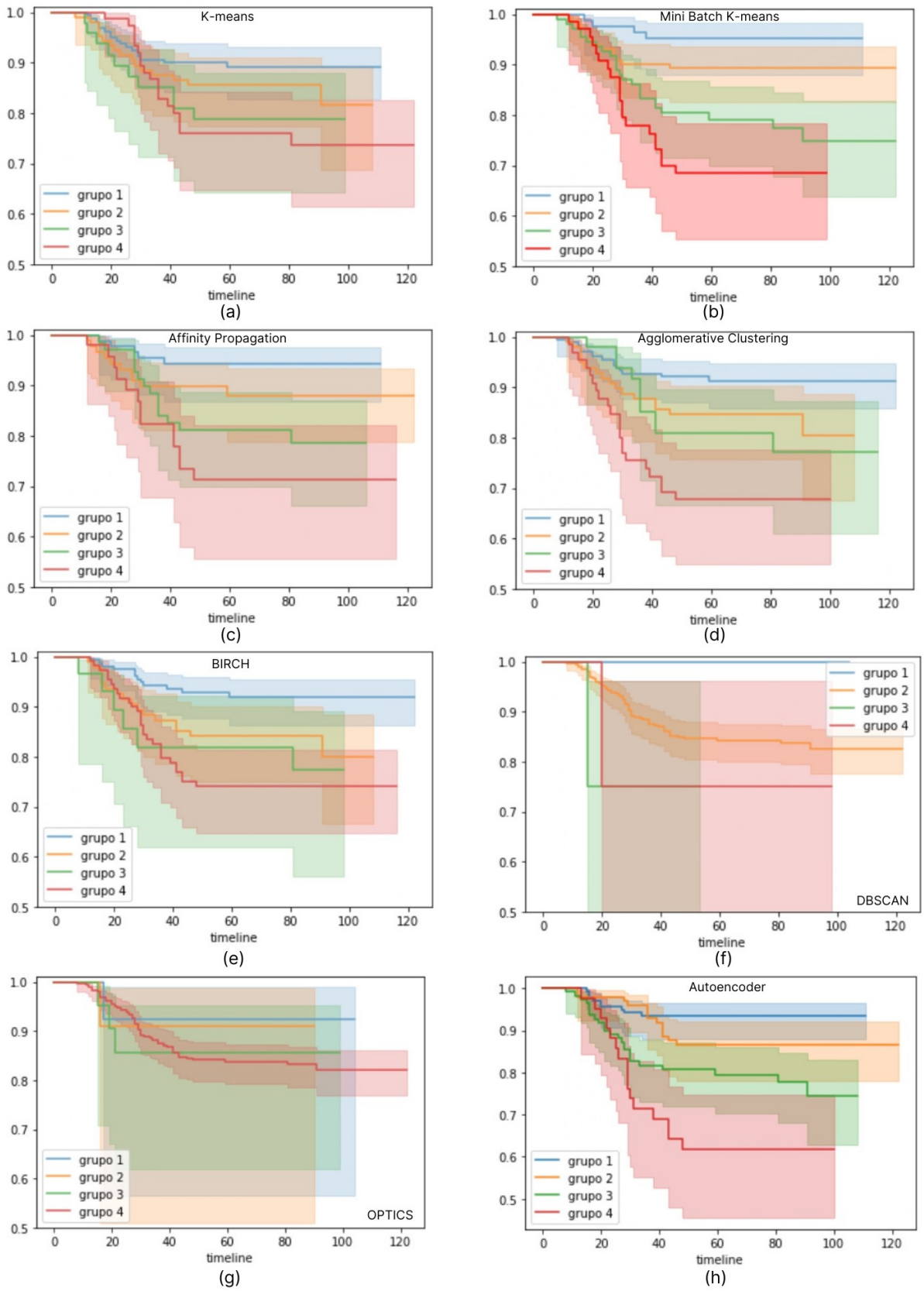
### 6.3.1 Quatro grupos de risco

Os resultados dos oito algoritmos de agrupamento para definir 4 grupos de risco podem ser vistos na Figura 22 e nas Tabelas 19 e 20. É possível observar que os grupos estão muito mais próximos do que nas primeiras estratégias, mas ainda assim é possível visualizar diferentes linhas nos gráficos. Novamente, os algoritmos de métodos hierárquicos DBSCAN e OPTICS (Figuras 21(f) e 21(g)) foram os que apresentaram os resultados menos satisfatórios. Os valores obtidos pelo teste log-rank nestes algoritmos foram altos (p-valor > 0,25), indicando que não há diferença significativa entre os grupos, e ambos os valores de coeficiente de silhueta foram negativos (-0,32156 e -0,13133, respectivamente).

Visualmente, os gráficos dos algoritmos Mini Batch K-means e Autoencoder foram os que apresentam grupos mais separados. Isso se confirma com o teste log-rank, em que os menores valores foram obtidos pelos dois algoritmos (p-valor < 0,2) e seus coeficientes de silhueta são positivos (0,13219 e 0,10901, respectivamente).

Ao observar os resultados para sugerir o algoritmo mais apropriado para a estratégia, foi escolhido o Mini Batch K-means. Embora os valores de teste de log-rank no algoritmo Autoencoder serem um pouco menor em comparação ao Mini Batch K-means, o coeficiente de silhueta do Autoencoder é o menor valor positivo obtido pelos algoritmos. Portanto, o Mini Batch K-means é o algoritmo mais apropriado para o objetivo que estamos buscando, pois apresenta todos os valores de teste log-rank menores que 0,20 e o coeficiente de silhueta positivo no valor de 0,13219.

Figura 22 – Gráficos de Kaplan-Meier dos agrupamentos resultantes para formar quatro grupos de risco, sendo os algoritmos (a) K-means, (b) Mini Batch K-means, (c) Affinity Propagation, (d) Agglomerative Clustering, (e) BIRCH, (f) DBSCAN, (g) OPTICS e (h) Autoencoder




Fonte: Elaborado pela autora.

Tabela 19 - p-valor obtido em cada algoritmo no teste log-rank

| | *p-valor* (teste log-rank) | | | | | |
|---|---|---|---|---|---|---|
| **Algoritmo** | **Grupos 1 e 2** | **Grupos 1 e 3** | **Grupos 1 e 4** | **Grupos 2 e 3** | **Grupos 2 e 4** | **Grupos 3 e 4** |
| K-means | 0,27 | **0,06** | **0,01** | 0,37 | 0,15 | 0,71 |
| Mini Batch K-Means | 0,18 | **<0,005** | **<0,005** | **<0,005** | **0,01** | 0,11 |
| Affinity Propagation | **<0,005** | **0,01** | **0,01** | 0,35 | 0,29 | 0,85 |
| Agglomerative Clustering | **0,05** | **0,02** | **<0,005** | 0,6 | **0,02** | 0,15 |
| BIRCH | **0,02** | **0,02** | **<0,005** | 0,57 | 0,14 | 0,74 |
| DBSCAN | 1 | 0,26 | 0,31 | 0,33 | 0,32 | 0,4 |
| OPTICS | 0,88 | 0,54 | 0,4 | 0,69 | 0,58 | 0,82 |
| Autoencoder | 0,09 | **<0,005** | **<0,005** | 0,09 | **<0,005** | **0,05** |

Fonte: Elaborado pela autora.

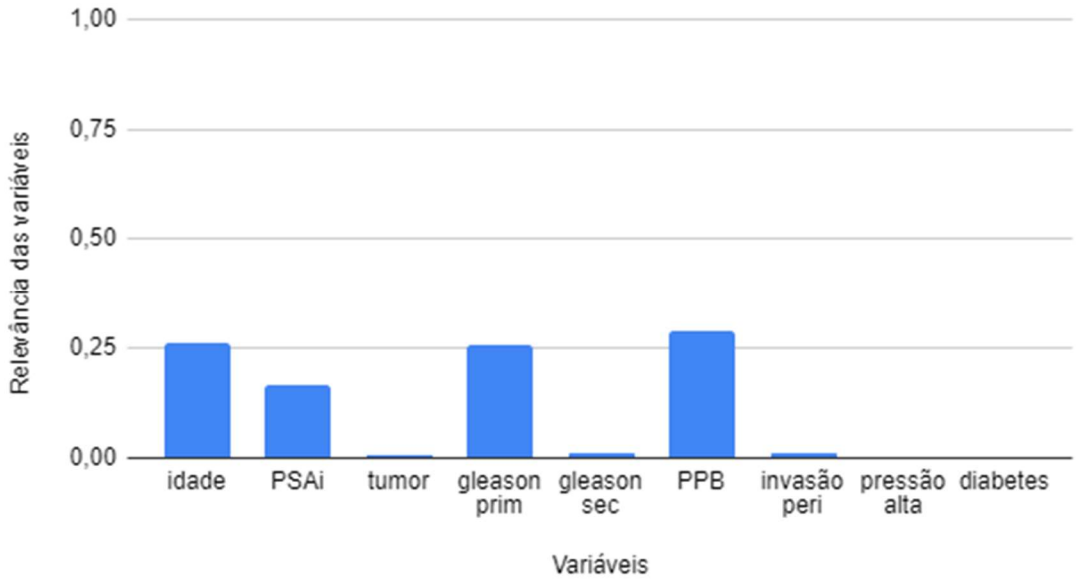
Tabela 20 - Valores de coeficiente de silhueta obtidos em cada algoritmo. Os únicos valores negativos foram obtidos pelos algoritmos DBSCAN e OPTICS.

| **Algoritmo** | **Coeficiente de Silhueta** |
|---|---|
| K-means | 0,17016 |
| Mini Batch K-Means | 0,13219 |
| Affinity Propagation | 0,15520 |
| Agglomerative Clustering | 0,13688 |
| BIRCH | 0,12551 |
| DBSCAN | -0,32156 |
| OPTICS | -0,13133 |
| Autoencoder | 0,10901 |

Fonte: Elaborado pela autora.

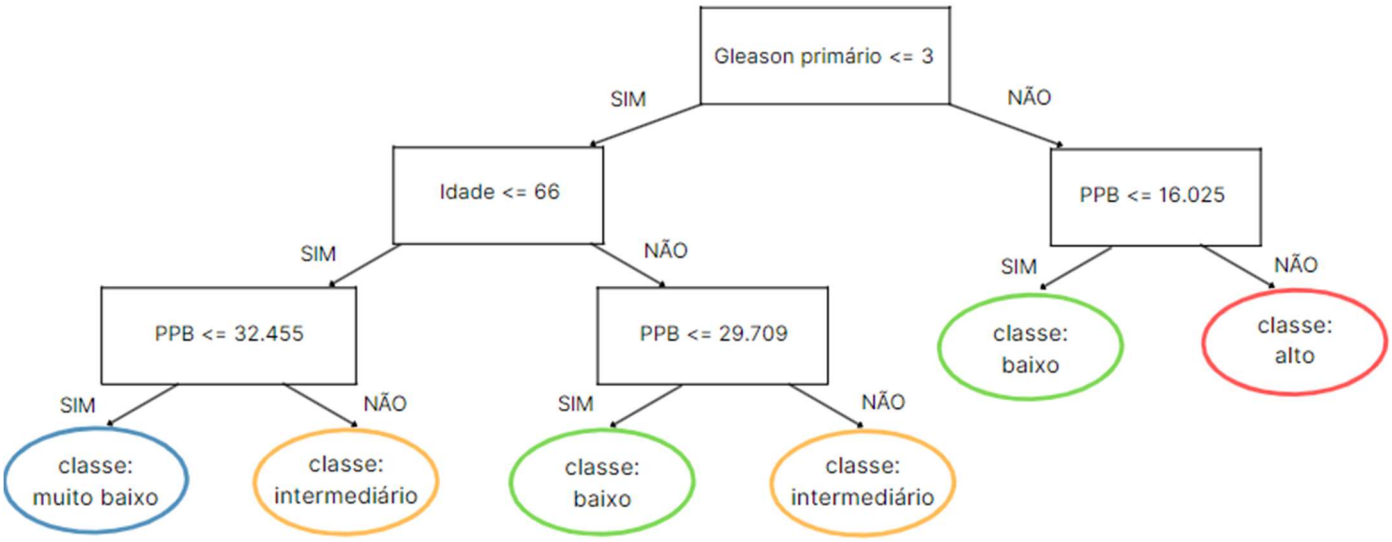
Na Figura 23, podemos observar o quão relevante cada variável é para o agrupamento. As variáveis PPB, escore de Gleason primário e idade se mostram mais relevantes em comparação ao restante das variáveis. Já na árvore de decisão (Figura 24), o primeiro valor a ser considerado para a classificação é o escore de Gleason primário: se o valor for menor ou igual a 3, o paciente não é classificado no grupo de risco muito alto; caso contrário, o paciente pode ser classificado nos grupos de risco intermediário ou muito alto.

Figura 23 – Gráfico com a relevância das variáveis para a divisão em quatro grupos de risco realizado pelo algoritmo Mini Batch K-means. As variáveis consideradas mais relevantes para o agrupamento foram PPB, idade e escore de Gleason primário.




Fonte: Elaborado pela autora.

Figura 24 – Árvore de decisão implementada a partir dos quatro grupos de risco formados pelo algoritmo Mini Batch K-means

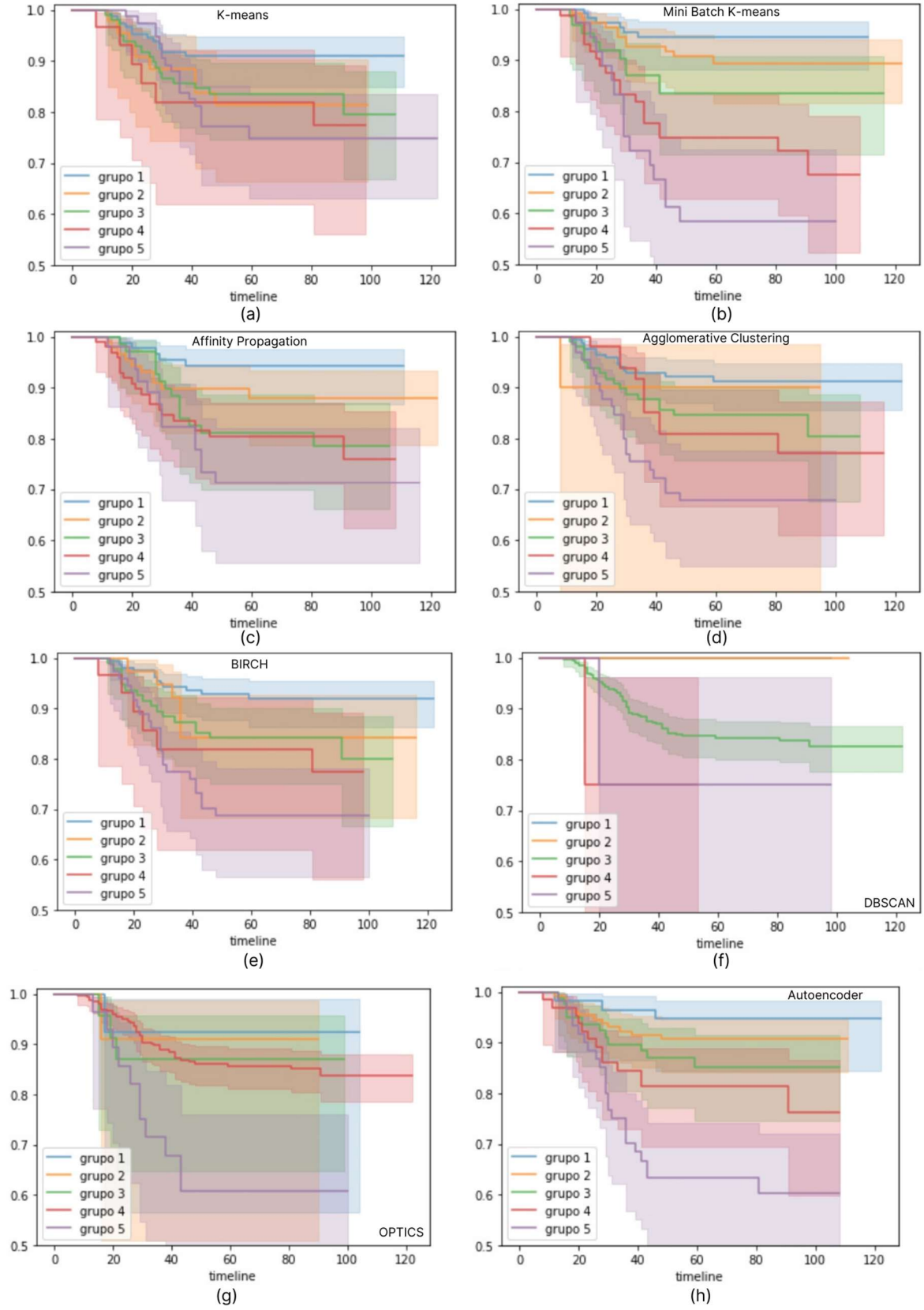



Fonte: Elaborado pela autora.

### 6.3.2 Cinco grupos de risco

Como os resultados de alguns métodos de agrupamento indicaram 4 grupos separados visualmente no gráfico Kaplan-Meier, conforme mostrado anteriormente, o próximo passo foi dividir os dados em 5 grupos no total. O resultado dos gráficos de Kaplan-Meier pode ser observado na Figura 25. Ao dividir os dados em 5 grupos, os algoritmos não obtiveram os resultados desejados. A maioria dos algoritmos não dividiu os grupos de risco de forma que as curvas que representam os grupos ficassem distantes visualmente no gráfico de Kaplan-Meier ou obtivessem valores baixos no teste de log-rank. Entre todos os algoritmos, novamente, os resultados mais satisfatórios foram obtidos pelos algoritmos Mini Batch K-means e Autoencoder. Os valores do teste de log-rank e do coeficiente de silhueta podem ser observados nas Tabelas 21 e 22, respectivamente.

Figura 25 – Gráficos de Kaplan-Meier dos agrupamentos resultantes para formar cinco grupos de risco, sendo os algoritmos (a) K-means, (b) Mini Batch K-means, (c) Affinity Propagation, (d) Agglomerative Clustering, (e) BIRCH, (f) DBSCAN, (g) OPTICS e (h) Autoencoder




Fonte: Elaborado pela autora.

Tabela 21 - p-valor obtido em cada algoritmo no teste log-rank, com 5 grupos

| | ***p-valor* (teste log-rank)** | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Algoritmo** | **Grupos 1 e 2** | **Grupos 1 e 3** | **Grupos 1 e 4** | **Grupos 1 e 5** | **Grupos 2 e 3** | **Grupos 2 e 4** | **Grupos 2 e 5** | **Grupos 3 e 4** | **Grupos 3 e 5** | **Grupos 4 e 5** |
| K-means | **0,08** | **0,06** | **0,05** | **<0,005** | 0,87 | 0,73 | 0,53 | 0,61 | 0,38 | 0,89 |
| Mini Batch K-Means | 0,20 | **0,02** | **<0,005** | **<0,005** | 0,22 | **<0,005** | **<0,005** | 0,13 | **0,01** | 0,17 |
| Affinity Propagation | 0,17 | **0,01** | **<0,005** | **<0,005** | 0,16 | **0,1** | **0,02** | 0,83 | 0,3 | 0,37 |
| Agglomerative Clustering | 0,83 | **0,05** | **0,02** | **<0,005** | 0,65 | 0,44 | 0,2 | 0,6 | **0,02** | 0,15 |
| BIRCH | 0,13 | **0,02** | **0,02** | **<0,005** | 0,82 | 0,51 | 0,08 | 0,57 | **0,03** | 0,41 |
| DBSCAN | 1 | 0,26 | 0,31 | 0,32 | 0,26 | 0,32 | 0,33 | 0,4 | 0,58 | 0,81 |
| OPTICS | 0,88 | 0,64 | 0,48 | **0,06** | 0,75 | 0,64 | **0,06** | 0,87 | **0,09** | **<0,005** |
| Autoencoder | 0,37 | 0,1 | **0,02** | **<0,005** | 0,28 | **0,04** | **<0,005** | 0,35 | **<0,005** | **0,03** |

Fonte: Elaborado pela autora.

Tabela 22 - Valores de coeficiente de silhueta obtidos em cada algoritmo, com 5 grupos. Além dos algoritmos DBSCAN e OPTICS, o algoritmo Affinity Propagation também obteve um valor negativo.

| **Algoritmo** | **Coeficiente de Silhueta** |
|---|---|
| K-means | 0,17357 |
| Mini Batch K-Means | 0,13633 |
| Affinity Propagation | -0,06595 |
| Agglomerative Clustering | 0,15277 |
| BIRCH | 0,13899 |
| DBSCAN | -0,46279 |
| OPTICS | -0,23154 |
| Autoencoder | 0,10059 |

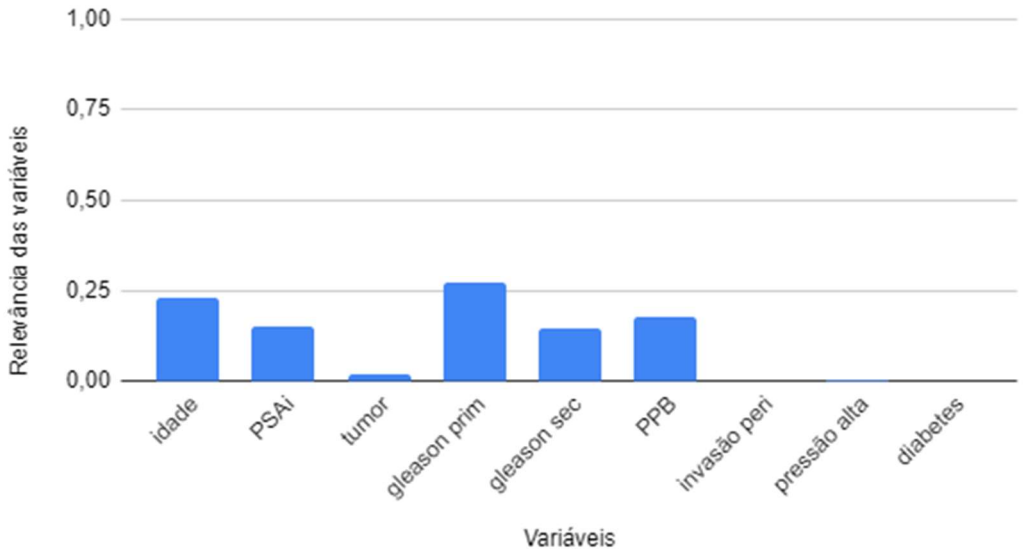
Fonte: Elaborado pela autora.

Ao analisar os valores do teste log-rank, assim como em outras estratégias, os resultados mais satisfatórios foram obtidos pelos algoritmos Mini Batch K-means e Autoencoder, ambos com valores baixos, indicando que na maioria das comparações os grupos formados apresentaram diferença de forma significativa. Em relação aos valores de coeficiente de silhueta, além dos algoritmos DBSCAN e OPTICS, o algoritmo Affinity Propagation também obteve um valor negativo (-0,06595), e o restante doa algoritmos obtiveram um valor positivo.

Portanto, comparando os resultados de todos os algoritmos, o algoritmo Mini Batch K-means foi o algoritmo que obteve os melhores resultados para a definição de 5 grupos de risco.

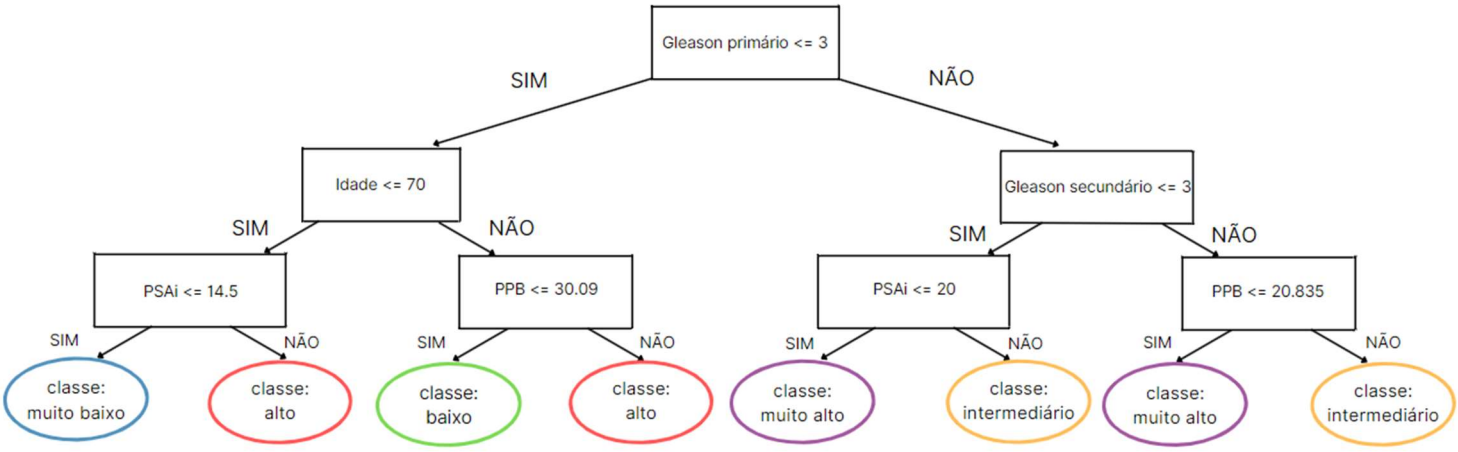
O gráfico na Figura 26 mostra um equilíbrio entre as relevâncias das variáveis para o agrupamento, com destaque para as variáveis escore de Gleason primário, idade, PPB, PSAi e escore de Gleason secundário. A árvore de decisão, ilustrada na Figura 27, revela novamente um viés nos dados. Uma vez que o valor de escore de Gleason primário é maior que 3, o paciente que tiver os dados com os maiores valores é classificado no grupo de risco intermediário e, se contrário, é classificado no grupo de risco muito alto. Como o objetivo é classificar pacientes em grupos de risco de recidiva bioquímica, é esperado que os maiores valores das variáveis indiquem um maior risco de recidiva bioquímica ao paciente, o que não ocorre neste agrupamento.

Figura 26 – Gráfico com a relevância das variáveis para a divisão em cinco grupos de risco realizado pelo algoritmo Mini Batch K-means, com destaque para o escore de Gleason primário, idade, PPB, PSAi e escore de Gleason secundário.




Fonte: Elaborado pela autora.

Figura 27 – Árvore de decisão construída a partir dos cinco grupos de risco formados pelo algoritmo Mini Batch K-means




Fonte: Elaborado pela autora.

Desse modo, apesar do algoritmo Mini Batch K-means apresentar resultados satisfatórios ao agrupar os dados em 5 diferentes grupos de risco, a análise de influência das variáveis mais relevantes no agrupamento para a classificação de pacientes revela um resultado não satisfatório. Portanto, esta divisão foi desconsiderada e não foram feitas demais análises incrementando o número de grupos de risco para um valor maior que 5.

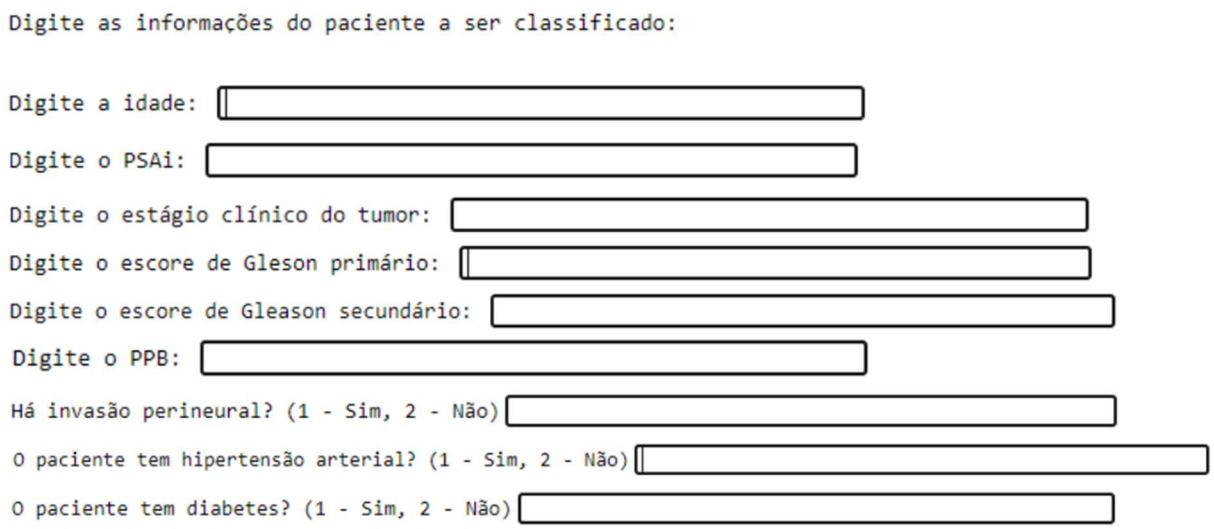
## 6.4 Ferramenta

Após realizar todas as análises dos resultados, uma ferramenta foi desenvolvida para que o usuário consiga classificar novos pacientes nos grupos de risco definidos pelos melhores resultados das três estratégias. A ferramenta foi desenvolvida com a linguagem de programação Python utilizando o ambiente de desenvolvimento Jupyter Notebook e o método predict, que recebe como parâmetro as variáveis do paciente e como resultado indica qual o grupo mais próximo que o paciente pertence. Como a subdivisão realizada a partir do grupo de risco alto do método de D'Amico não obteve resultados satisfatórios, a ferramenta não considera os subgrupos alto e muito alto ao classificar o paciente em subgrupos dos grupos de risco D'Amico.

Na ferramenta, o usuário deve informar os valores das 9 variáveis utilizadas para a classificação, digitando os valores nas caixas de texto ao lado do nome de cada variável (Figura 28). Após o preenchimento dos campos, a ferramenta classifica o novo paciente em um grupo

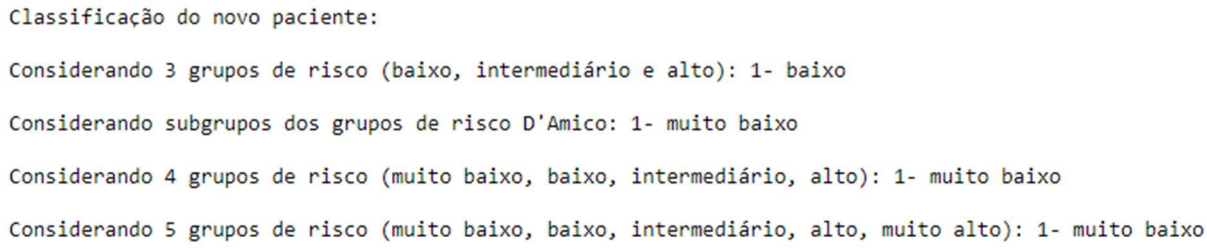
de risco mais adequado nas três estratégias diferentes e baseado nos melhores resultados discutidos neste capítulo (Figura 29). Assim, o usuário pode analisar os grupos de riscos em que o novo paciente foi classificado e escolher a estratégia que achar mais adequada.

Figura 28 – Ferramenta desenvolvida para classificação de um novo paciente e seus campos a serem preenchidos.




Fonte: Elaborado pela autora.

Figura 29 – Exemplo de resultado da ferramenta.




Fonte: Elaborado pela autora.

# 7 CONCLUSÕES

Este trabalho teve como objetivo obter um refinamento dos grupos de risco em pacientes com câncer de próstata tratados com radioterapia, utilizando técnicas não supervisionadas de Aprendizado de Máquina tradicionais e de Aprendizagem Profunda, para geração de agrupamentos, a partir de três diferentes abordagens propostas. Além disso, também foram objetivos deste trabalho identificar indicadores relevantes para o melhor refinamento dos grupos; estabelecer diferentes estratégias de obtenção dos agrupamentos e avaliar pontos fortes e fracos de cada uma; e comparar os diferentes algoritmos de aprendizado não supervisionado.

Para que estes objetivos fossem alcançados, dados de 404 pacientes com câncer de próstata tratados com radioterapia foram utilizados na implementação de oito diferentes algoritmos de agrupamento. O gráfico de Kaplan-Meier, o teste log-rank e o método de silhueta foram utilizados para avaliar e analisar os agrupamentos resultantes de cada algoritmo. A árvore de decisão foi implementada a partir dos indicadores mais relevantes identificados pelos melhores resultados dos algoritmos para apresentar aos usuários conjuntos de regras que possibilita entender a lógica de geração dos agrupamentos.

Dentre os algoritmos implementados, os algoritmos baseados em densidade (DBSCAN e OPTICS) apresentaram os resultados menos satisfatórios. Os gráficos de Kaplan-Meier resultantes não apresentavam curvas bem separadas, os valores do teste log-rank eram altos e os valores de coeficiente de silhueta, em sua maioria, negativos.

Considerando as três estratégias propostas, os algoritmos em destaque com os melhores resultados foram Mini Batch K-means e Autoencoder. Ambos os algoritmos apresentaram gráficos de Kaplan-Meier com curvas bem definidas e distintas, valores baixos obtidos a partir do teste log-rank e coeficientes de silhueta positivos. Para as Estratégias 1 e 3, o algoritmo Mini Batch K-means foi considerado o mais apropriado. O algoritmo Autoencoder foi considerado o mais apropriado para definir todos os subgrupos da Estratégia 2.

Ao implementar a árvore de decisão a partir dos resultados desses algoritmos, foi observado que é possível obter regras de agrupamentos refinadas seguindo estas estratégias. Para a divisão dos pacientes em 3 e 4 grupos de risco (Estratégias 1 e 3, respectivamente) e para subdividir o grupo de risco intermediário em 2 e 3 subgrupos (Estratégia 2), é possível observar a importância do escore de Gleason primário. Em todos os casos citados, esse indicador é o primeiro considerado na árvore de decisão. Além do mais, o escore de Gleason primário é o único considerado pela árvore de decisão implementada ao subdividir o grupo de risco

intermediário em 2 subgrupos. Como nas regras de agrupamento do método de D'Amico e em diversos trabalhos correlatos, este indicador não é considerado de forma independente, pois é somado ao escore de Gleason secundário para obter a soma de escore de Gleason total, é importante obter a informação de como pode influenciar no agrupamento.

De forma geral, ficou destacada a relevância das variáveis escore de Gleason primário, idade, PPB, escore de Gleason secundário e estágio clínico do tumor. Destas, escore de Gleason primário, idade e PPB se sobressaem. As variáveis escore de Gleason primário, escore de Gleason secundário e estágio clínico do tumor também estão presentes nas regras de classificação do grupo de risco de D'Amico, diferente das variáveis idade e PPB. Já a variável PSAi, que faz parte das regras de D'Amico, não se destacou como uma variável relevante nos agrupamentos realizados pelos algoritmos de Aprendizado de Máquina. É possível que em uma coorte maior, com mais dados de pacientes, esta relevância seja revelada/observada/visível.

Com isso, uma das limitações deste estudo é a quantidade de dados. Com uma quantidade maior, e mais pacientes, pode ser que seja possível obter resultados mais consistentes. Outra limitação é a ausência de "não-grupo", pois os algoritmos utilizados classificaram todos os pacientes em um dos grupos obtidos.

O presente estudo apresenta uma abordagem de definição de grupos de risco com vantagens em relação a abordagem amplamente utilizada seguindo as regras de D'Amico. As estratégias propostas obtiveram uma divisão mais acurada por definirem um número maior de grupos de risco, que apresentam distinção entre si e consideram variáveis que não são considerados nas regras de D'Amico. Além disso, essas variáveis tiveram relevância destacada nos agrupamentos realizados pelos algoritmos de Aprendizado de Máquina, indicando que o presente estudo mostra que esses indicadores são relevantes para a predição do grupo de risco do paciente.

Por fim, com a implementação da ferramenta considerando diferentes estratégias e algoritmos apresentados, conclui-se que este estudo apresenta novas regras de grupos de risco refinadas, baseadas em dados não somente referentes ao tumor, mas também do próprio paciente, além de oferecer uma maneira mais simples e direta para a compreensão do médico especialista a respeito das regras intrínsecas a cada abordagem de classificação, auxiliando na tomada de decisão em relação ao planejamento do tratamento do paciente.

Como trabalhos futuros, os resultados poderão vir a ser analisados por um especialista em radioterapia, com experiência com pacientes com câncer de próstata, para avaliar os grupos formados pelos algoritmos e as características dos pacientes pertencentes a cada grupo. Além disso, a base de dados pode ser incrementada continuamente com dados de mais pacientes, os

agrupamentos e as estratégias podem ser remodelados e, com isso, aprimorar a precisão e a sensibilidade dos grupos de risco.

Para melhor experiência do usuário, também como trabalho futuro, será desenvolvida a interface da ferramenta implementada para determinação do grupo de risco de um novo paciente. Também será implementada a funcionalidade da ferramenta aceitar um arquivo com informações de diversos pacientes e classificar todos em um único processamento.

# REFERÊNCIAS

ACKERMANN, Marcel R. et al. Analysis of agglomerative clustering. Algorithmica, v. 69, n. 1, p. 184-215, 2014.

AMARBAYASGALAN, Tsatsral; JARGALSAIKHAN, Bilguun; RYU, Keun Ho. Unsupervised novelty detection using deep autoencoders with density based clustering. Applied Sciences, v. 8, n. 9, p. 1468, 2018.

ANKERST, Mihael et al. OPTICS: Ordering points to identify the clustering structure. ACM Sigmod record, v. 28, n. 2, p. 49-60, 1999.

ARVANITI, Eirini et al. Automated Gleason grading of prostate cancer tissue microarrays via deep learning. Scientific reports, v. 8, n. 1, p. 1-11, 2018.

BAYNE, Christopher E.; JARRETT, Thomas W. Cancer of the prostate: incidence in the USA. In: Prostate Cancer. Academic Press, 2016. p. 119-125.

BÉJAR ALONSO, Javier. K-means vs mini batch k-means: a comparison. 2013.

BERLIN, Alejandro et al. International multicenter validation of an intermediate risk subclassification of prostate cancer managed with radical treatment without hormone therapy. The Journal of urology, v. 201, n. 2, p. 284-291, 2019.

BEWICK, Viv; CHEEK, Liz; BALL, Jonathan. Statistics review 12: survival analysis. Critical care, v. 8, n. 5, p. 1-6, 2004.

BREIMAN, Leo et al. Classification and regression trees. Routledge, 1984.

BRIJAIN, Mr et al. A survey on decision tree algorithm for classification. 2014.

VAN BOOVEN, Derek J. et al. A systematic review of artificial intelligence in prostate cancer. Research and reports in urology, v. 13, p. 31, 2021.

CHARIKAR, Moses et al. Incremental clustering and dynamic information retrieval. SIAM Journal on Computing, v. 33, n. 6, p. 1417-1440, 2004.

CHOLLET, Francois. Deep learning with Python. Simon and Schuster, 2017.

CHOLLET, François et al. keras. 2015.

CHURILOV, Leonid et al. Improving risk grouping rules for prostate cancer patients with optimization. In: 37th Annual Hawaii International Conference on System Sciences, 2004. Proceedings of the. IEEE, 2004. p. 9 pp.

CHURILOV, Leonid et al. Data mining with combined use of optimization techniques and self-organizing maps for improving risk grouping rules: application to prostate cancer patients. Journal of Management Information Systems, v. 21, n. 4, p. 85-100, 2005.

CRUZ, Joseph A.; WISHART, David S. Applications of machine learning in cancer prediction and prognosis. Cancer informatics, v. 2, p. 59, 2006.

D'AMICO, Anthony V. et al. Pretreatment nomogram for prostate-specific antigen recurrence after radical prostatectomy or external-beam radiation therapy for clinically localized prostate cancer. Journal of Clinical Oncology, v. 17, n. 1, p. 168-168, 1999.

DAVIDSON-PILON, Cameron. lifelines: survival analysis in Python. Journal of Open Source Software, v. 4, n. 40, p. 1317, 2019.

EMINAGA, Okyaz et al. Combination possibility and deep learning model as clinical decision-aided approach for prostate cancer. Health informatics journal, v. 26, n. 2, p. 945-962, 2020.

ESTER, Martin et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In: kdd. 1996. p. 226-231.

FREY, Brendan J.; DUECK, Delbert. Clustering by passing messages between data points. science, v. 315, n. 5814, p. 972-976, 2007.

GNANAPRAGASAM, Vincent J. et al. Improving clinical risk stratification at diagnosis in primary prostate cancer: a prognostic modelling study. PLoS medicine, v. 13, n. 8, p. e1002063, 2016.

GOEL, Manish Kumar; KHANNA, Pardeep; KISHORE, Jugal. Understanding survival analysis: Kaplan-Meier estimate. International journal of Ayurveda research, v. 1, n. 4, p. 274, 2010.

HAN, Jiawei; KAMBER, Micheline; PEI, Jian. 10-cluster analysis: Basic concepts and methods. In: Data mining. Morgan Kaufmann, 2012. p. 443-495.

HAO, Xing; ZHANG, Guigang; MA, Shang. Deep learning. International Journal of Semantic Computing, v. 10, n. 03, p. 417-439, 2016.

HARRIS, Charles R. et al. Array programming with NumPy. Nature, v. 585, n. 7825, p. 357-362, 2020.

HERNANDEZ, David J. et al. Contemporary evaluation of the D'amico risk classification of prostate cancer. Urology, v. 70, n. 5, p. 931-935, 2007.
HENNIG, Christian et al. (Ed.). Handbook of cluster analysis. CRC Press, 2015.

HUNTER, John D. Matplotlib: A 2D graphics environment. Computing in science & engineering, v. 9, n. 03, p. 90-95, 2007.

INSTITUTO NACIONAL DO CÂNCER. Síntese de Resultados e Comentários. Câncer de Próstata. Rio de Janeiro: INCA, 2020. Disponível em: https://www.inca.gov.br/estimativa/sintese-de-resultados-e-comentarios. Acesso em: 15 junho 2021.

JAGER, Kitty J. et al. The analysis of survival data: the Kaplan–Meier method. Kidney international, v. 74, n. 5, p. 560-565, 2008.

JAIN, Anil K. Data clustering: 50 years beyond K-means. Pattern recognition letters, v. 31, n. 8, p. 651-666, 2010.

JAIN, Anil K.; MURTY, M. Narasimha; FLYNN, Patrick J. Data clustering: a review. ACM computing surveys (CSUR), v. 31, n. 3, p. 264-323, 1999.

KHAN, Kamran et al. DBSCAN: Past, present and future. In: The fifth international conference on the applications of digital information and web technologies (ICADIWT 2014). IEEE, 2014. p. 232-238.

KUMAR, Neeraj et al. Convolutional neural networks for prostate cancer recurrence prediction. In: Medical Imaging 2017: Digital Pathology. International Society for Optics and Photonics, 2017. p. 101400H.

MCKINNEY, Wes et al. Data structures for statistical computing in python. In: Proceedings of the 9th Python in Science Conference. 2010. p. 51-56.

MIN, Erxue et al. A survey of clustering with deep learning: From the perspective of network architecture. IEEE Access, v. 6, p. 39501-39514, 2018.

MOHLER, James L. et al. Prostate cancer, version 2.2014. Journal of the National Comprehensive Cancer Network, v. 12, n. 5, p. 686-718, 2014.

MYLES, Anthony J. et al. An introduction to decision tree modeling. Journal of Chemometrics: A Journal of the Chemometrics Society, v. 18, n. 6, p. 275-285, 2004.

NATIONAL COMPREHENSIVE CANCER NETWORK [NCCN]. NCCN Guidelines for Patients Advanced Stage Prostate Cancer (2020). Disponível em: https://www.nccn.org/patients/guidelines/content/PDF/prostate-advanced-patient.pdf. Acesso em 1 de out. de 2021.

NELSON, William G. et al. Prostate Cancer. Abeloff's Clinical Oncology, p. 1401–1432. e6, 2020.

NIELSEN, Frank. Introduction to HPC with MPI for Data Science. Springer, 2016.

PEDREGOSA, Fabian et al. Scikit-learn: Machine learning in Python. the Journal of machine Learning research, v. 12, p. 2825-2830, 2011.

RAFSANJANI, M. Kuchaki; VARZANEH, Z. Asghari; CHUKANLO, N. Emami. A survey of hierarchical clustering algorithms. The Journal of Mathematics and Computer Science, v. 5, n. 3, p. 229-240, 2012.

REESE, Adam C. et al. Contemporary evaluation of the National Comprehensive Cancer Network prostate cancer risk classification system. Urology, v. 80, n. 5, p. 1075-1079, 2012.

ROKACH, Lior; MAIMON, Oded. Clustering methods. In: Data mining and knowledge discovery handbook. Springer, Boston, MA, 2005. p. 321-352.

ROSS, Phillip L. et al. Comparisons of nomograms and urologists' predictions in prostate cancer. In: Seminars in urologic oncology. 2002. p. 82-88.

ROUSSEEUW, Peter J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of computational and applied mathematics, v. 20, p. 53-65, 1987.

SALEKSHAHREZAEE, Zahra; LEEVY, Joffrey L.; KHOSHGOFTAAR, Taghi M. A reconstruction error-based framework for label noise detection. Journal of Big Data, v. 8, n. 1, p. 1-16, 2021.

SANDA, Martin G. et al. Clinically localized prostate cancer: AUA/ASTRO/SUO guideline. Part I: risk stratification, shared decision making, and care options. The Journal of urology, v. 199, n. 3, p. 683-690, 2018.

SASIREKHA, K.; BABY, P. Agglomerative hierarchical clustering algorithm-a. International Journal of Scientific and Research Publications, v. 83, p. 83, 2013.

SCULLEY, David. Web-scale k-means clustering. In: Proceedings of the 19th international conference on World wide web. 2010. p. 1177-1178.

SHUKLA, Nagesh et al. Breast cancer data analysis for survivability studies and prediction. Computer methods and programs in biomedicine, v. 155, p. 199-208, 2018.

SUAREZ-IBARROLA, Rodrigo et al. Current and future applications of machine and deep learning in urology: a review of the literature on urolithiasis, renal cell carcinoma, and bladder and prostate cancer. World journal of urology, v. 38, n. 10, p. 2329-2347, 2020.

THEODORIDIS, Sergios; KOUTROUMBAS, Konstantinos. Pattern Recognition. Asterdam: Elvesier, 2009.

WARD JR, Joe H. Hierarchical grouping to optimize an objective function. Journal of the American statistical association, v. 58, n. 301, p. 236-244, 1963.

YAO, Xin; LIU, Yong. Machine learning. In: Search Methodologies. Springer, Boston, MA, 2014. p. 477-517.

ZHANG, Xian-Da. Machine learning. In: A Matrix Algebra Approach to Artificial Intelligence. Springer, Singapore, 2020. p. 223-440.

ZHANG, Guijuan; LIU, Yang; JIN, Xiaoning. A survey of autoencoder-based recommender systems. Frontiers of Computer Science, v. 14, n. 2, p. 430-450, 2020.

ZUMSTEG, Zachary S. et al. A new risk classification system for therapeutic decision making with intermediate-risk prostate cancer patients undergoing dose-escalated external-beam radiation therapy. European urology, v. 64, n. 6, p. 895-902, 2013.

ZUPAN, Blaž et al. Machine learning for survival analysis: a case study on recurrence of prostate cancer. Artificial intelligence in medicine, v. 20, n. 1, p. 59-75, 2000.

# ANEXO

## ANEXO A – Parecer do Comitê de Ética e Pesquisa do HC-FMRP

USP - HOSPITAL DAS CLÍNICAS DA FACULDADE DE MEDICINA DE RIBEIRÃO PRETO DA USP -

### PARECER CONSUBSTANCIADO DO CEP

**DADOS DO PROJETO DE PESQUISA**

**Título da Pesquisa:** Utilização de aprendizado de máquina para refinamento do grupo de risco de câncer de próstata em pacientes tratados com radioterapia

**Pesquisador:** Thaine Cravo Marques dos Santos

**Área Temática:**

**Versão:** 1

**CAAE:** 51052821.9.0000.5440

**Instituição Proponente:** Hospital das Clínicas da Faculdade de Medicina de Ribeirão Preto da USP -

**Patrocinador Principal:** Financiamento Próprio

**DADOS DO PARECER**

**Número do Parecer:** 4.943.463

**Apresentação do Projeto:**

O câncer de próstata é o segundo câncer mais frequente em homens no mundo. A radioterapia apresenta um papel fundamental no tratamento do câncer de próstata. Para planejamento do tratamento inicial, o paciente é classificado em um grupo de risco, podendo ser baixo, intermediário ou alto risco. Entretanto, esses grupos apresentam grande heterogeneidade, e assim pacientes com diferentes características são classificados em um mesmo grupo de risco. O tratamento excessivo de tumores com pouca probabilidade de progressão e o subtratamento de tumores mais agressivos são consequências disso. As técnicas de aprendizado de máquina vêm sendo utilizadas em diversas aplicações em câncer, incluindo classificação e avaliação de risco. Com isso, o objetivo deste projeto é refinar o grupo de risco de câncer de próstata em pacientes tratados com radioterapia utilizando métodos de aprendizado de máquina não supervisionados a fim de formar grupos de risco mais homogêneos e melhorar a precisão do risco de falha bioquímica. Serão utilizados dados de 485 pacientes com câncer de próstata tratados com radioterapia entre janeiro de 2010 e janeiro de 2017 no Hospital das Clínicas da Faculdade de Medicina de Ribeirão Preto. Os critérios de inclusão são: possuir idade entre 44 e 96 anos no momento do diagnóstico e utilização de Radioterapia Conformada 3D (3D-CRT) ou Radioterapia de Intensidade Modulada (IMRT) com uma dose total maior ou igual a 74 Gy. Os critérios de exclusão são: pacientes com metástases, história prévia de prostatectomia, tratamento quimioterápico ou

**Endereço:** CAMPUS UNIVERSITÁRIO
**Bairro:** MONTE ALEGRE **CEP:** 14.048-900
**UF:** SP **Município:** RIBEIRAO PRETO
**Telefone:** (16)3602-2228 **Fax:** (16)3633-1144 **E-mail:** cep@hcrp.usp.br

tratados com radiação pélvica devido ao câncer de próstata e pacientes submetidos à radioterapia hipofracionada. A coleta dos dados será realizada junto ao Serviço de Radioterapia do Hospital das Clínicas da Faculdade de Medicina de Ribeirão Preto, sob a supervisão do Dr.Gustavo Viani Arruda, médico radioterapeuta do Departamento de Radioterapia. Para realizar o refinamento do grupo de risco, será feito o pré-processamento dos dados. Diferentes métodos de normalização serão testados e utilizados durante os estudos: Implementação dos algoritmos de pré-processamento de dados; Implementação dos algoritmos de agrupamento; Implementação do método Kaplan-Meier e teste log-rank.

**Objetivo da Pesquisa:**

Objetivo geral deste projeto é refinar o grupo de risco em pacientes com câncer de próstata tratados com radioterapia utilizando técnicas não supervisionadas de aprendizado de máquina, realizando diferentes estudos com os dados.

Objetivos específicos são: Comparar os diferentes algoritmos de aprendizado não supervisionado; Implementar uma ferramenta de predição do grupo de risco utilizando aprendizado supervisionado com os grupos de risco resultantes do estudo.

**Avaliação dos Riscos e Benefícios:**

Os pesquisadores mencionam que não há riscos ou prejuízos aos pacientes. Referem que utilizarão dados clínicos já disponíveis, sem intervenções clínicas e sem alterações ou influências na rotina ou tratamento dos participantes de pesquisa.

-

Os benefícios deste estudo incluem uma nova classificação de risco para pacientes com câncer de próstata, auxiliando a equipe médica no planejamento do tratamento de forma mais eficaz e melhorando a precisão da análise de sobrevida de cada grupo de risco.

**Comentários e Considerações sobre a Pesquisa:**

A pesquisa está bastante fundamentada. Os pesquisadores solicitam dispensa do TCLE em função do estudo ser retrospectivo, que utilizará apenas informações clínicas disponíveis na instituição sem utilização de material biológico.

**Considerações sobre os Termos de apresentação obrigatória:**

Todos os documentos foram adequadamente apresentados. Solicita a dispensa de aplicação do TCLE.

A dispensa do uso de TCLE se fundamenta por ser um estudo retrospectivo, que

utilizará apenas informações clínicas disponíveis na instituição sem utilização de

material biológico. Todos os dados serão manuseados e analisados de forma anônima, sem identificação dos participantes de pesquisa, e os resultados decorrentes do estudo serão apresentados de forma que não permite a identificação individual dos participantes. Além disso, o estudo será sem intervenções clínicas e sem alterações ou influências na rotina ou tratamento do participante de pesquisa, e consequentemente sem adição de riscos ou prejuízos ao bem-estar dos pacientes.

**Recomendações:**

não se aplica

**Conclusões ou Pendências e Lista de Inadequações:**

Diante do exposto e à luz da Resolução CNS 466/2012, o projeto de pesquisa, assim como a solicitação de dispensa de aplicação do Termo de Consentimento Livre e Esclarecido, podem ser enquadrados na categoria APROVADO.

**Considerações Finais a critério do CEP:**

Projeto Aprovado: Tendo em vista a legislação vigente, devem ser encaminhados ao CEP, relatórios parciais anuais referentes ao andamento da pesquisa e relatório final ao término do trabalho. Qualquer modificação do projeto original deve ser apresentada a este CEP em nova versão, de forma objetiva e com justificativas, para nova apreciação.

**Este parecer foi elaborado baseado nos documentos abaixo relacionados:**

| Tipo Documento | Arquivo | Postagem | Autor | Situação |
|---|---|---|---|---|
| Informações Básicas do Projeto | PB_INFORMAÇÕES_BÁSICAS_DO_PROJETO_1809802.pdf | 23/08/2021 12:36:33 | | Aceito |
| Projeto Detalhado / Brochura Investigador | projeto_Thaine_CEP_corrigido.pdf | 23/08/2021 12:36:04 | Thaine Cravo Marques dos Santos | Aceito |
| Outros | CRONOGRAMA_PROJETO_THAINE.pdf | 17/08/2021 11:42:14 | Thaine Cravo Marques dos Santos | Aceito |
| Outros | ORCAMENTO_FINANCEIRO_THAINE.pdf | 17/08/2021 11:41:38 | Thaine Cravo Marques dos Santos | Aceito |
| Outros | APROVACAO_UPC_ASS.jpg | 17/08/2021 11:39:28 | Thaine Cravo Marques dos Santos | Aceito |
| TCLE / Termos de Assentimento / | DISPENSA_TCLE.pdf | 17/08/2021 11:37:29 | Thaine Cravo Marques dos | Aceito |

| | | | | |
|---|---|---|---|---|
| Justificativa de Ausência | DISPENSA_TCLE.pdf | 17/08/2021 11:37:29 | Santos | Aceito |
| Folha de Rosto | FOLHA_ROSTO_ASS.pdf | 17/08/2021 11:28:26 | Thaine Cravo Marques dos Santos | Aceito |

**Situação do Parecer:**
Aprovado

**Necessita Apreciação da CONEP:**
Não

RIBEIRAO PRETO, 31 de Agosto de 2021

**Assinado por:**
**MARCIA GUIMARÃES VILLANOVA**
**(Coordenador(a))**