

Tiago Beltrão Lacerda

# **Utilização de Aprendizagem de Máquina na Classificação de Risco de Pacientes Oncológicos**

Recife

2020

Tiago Beltrão Lacerda

## **Utilização de Aprendizagem de Máquina na Classificação de Risco de Pacientes Oncológicos**

Dissertação apresentada ao programa de Mestrado em Engenharia de Software da CESAR.SCHOOL, como requisito para a obtenção do título de Mestre em Engenharia de Software.

**CESAR.SCHOOL**

Mestrado Profissional

Orientador: Ana Paula Carvalho Cavalcanti Furtado, D.Sc

Coorientador: Dra. Jurema Telles de Oliveira Lima, D.Sc

Recife

2020

Tiago Beltrão Lacerda

Utilização de Aprendizagem de Máquina na Classificação de Risco de Pacientes Oncológicos/ Tiago Beltrão Lacerda. – Recife, 2020-

154 p. : il. (algumas color.) ; 30 cm.

Orientador: Ana Paula Carvalho Cavalcanti Furtado, D.Sc

Dissertação – **CESAR.SCHOOL**

Mestrado Profissional, 2020.

1. Aprendizagem de Máquina. 2. Câncer. 2. AGA. I. Prof.<sup>a</sup> Ana Paula Carvalho Cavalcanti Furtado. II. CESAR.SCHOOL III. Utilização de Aprendizagem de Máquina na Classificação de Risco de Pacientes Oncológicos.

# ERRATA

Tiago Beltrão Lacerda

## **Utilização de Aprendizagem de Máquina na Classificação de Risco de Pacientes Oncológicos**

Dissertação apresentada ao programa de Mestrado em Engenharia de Software da CESAR.SCHOOL, como requisito para a obtenção do título de Mestre em Engenharia de Software.

Trabalho preliminar. Recife, 12 de junho de 2020:

---

**Ana Paula Carvalho Cavalcanti  
Furtado, D.Sc**  
Orientadora

---

**Rodrigo Carneiro Leão Vieira da  
Cunha, D.Sc**  
Avaliador interno

---

**Rafael Ferreira Leite de Mello, D.Sc**  
Avaliador Externo

Recife  
2020

*À minha filha, Juliana.*

# AGRADECIMENTOS

**Agradeço** à minha esposa, Kátia. Sem ela não começaria (e nem terminaria) este mestrado.

À minha orientadora, Prof<sup>a</sup>. Dra. **Ana Paula Furtado**, pela oportunidade de ser seu aluno e por toda dedicação empregada.

À minha co-orientadora, Dra. **Jurema Telles**, por ter nos recebido tão bem no IMIP e orientado nosso trabalho no campo da medicina.

Aos amigos do mestrado, **Alberto Medeiros** e **Regis Perez**.

À equipe de apoio do IMIP, especialmente **Gerlane**, **Diogo** e **Débora**, por nos receberem sempre tão bem.

*“Se dois homens vêm andando por uma estrada, cada um com um pão, e, ao se encontrarem, trocarem os pães, cada um vai embora com um. Se dois homens vêm andando por uma estrada, cada um com uma ideia, e, ao se encontrarem, trocarem as ideias, cada um vai embora com duas.*

*(Provérbio Chinês)*

# RESUMO

**Contexto:** Aprendizagem de máquina (AM) está sendo utilizada com sucesso em várias áreas da ciência, a medicina não é exceção. O câncer, sendo uma doença heterogênea, consistindo de vários subtipos e possíveis tratamentos, gera um grande volume de dados (*big data*). O IMIP - Instituto de Medicina Integral Professor Fernando Figueira é uma entidade filantrópica que realiza mais de 3000 atendimentos oncológicos por mês, tem como prática, realizar a Avaliação Geriátrica Ampla (AGA) dos pacientes oncológicos, avaliação que resulta em mais de 100 variáveis para cada paciente, em seis dimensões (atributos físicos, fatores sociais e ambientais, funcionais, mobilidade, psicológicos e medicamentos). **Objetivo:** explorar, por meio de AM, a capacidade preditiva presente na AGA para classificar os pacientes oncológicos em alto e baixo risco de óbito em 6 meses. **Método:** foi realizada a pesquisa utilizando *Design Science Research* (DSR), onde concluiu-se dois ciclos de design para desenvolvimento e otimização do modelo de AM mais adequado ao contexto. **Resultado:** desenvolveu-se um modelo de AM utilizando o algoritmo SVM que obteve uma *AUC Precision-Recall* igual a 0,660 e *AUC ROC* igual a 0,836. Além disso, identificou-se um conjunto mínimo de variáveis com capacidade preditiva superior ao conjunto completo. **Conclusão:** demonstrou-se, utilizando uma abordagem supervisionada de aprendizado de máquina em conjunto com a AGA, que é possível realizar a classificação de risco de pacientes oncológicos, além de otimizar o modelo de AM desenvolvido para melhor identificar os pacientes de alto risco.

**Palavras-chave:** Aprendizagem de Máquina. Câncer. AGA.

# ABSTRACT

**Context:** Machine Learning (ML) is being used successfully in many areas of science, medicine is no exception. Cancer, being a heterogeneous disease, consisting of various subtypes and possible treatments, generates a large volume of data (big data). IMIP - Instituto de Medicina Integral Professor Fernando Figueira, as a philanthropic entity that performs more than 3000 oncologic attendances per month, performs a Comprehensive Geriatric Assessment (CGA) on oncologic patients, which results in more than 100 variables for each patient in six dimensions (physical attributes, social and environmental factors, functional, mobility, psychological and medication). **Objective:** to explore, by using ML, the predictive capacity present in the CGA to classify oncologic patients in high and low risk of death in 6 months. **Method:** the research was conducted by using the Design Science Research (DSR), where two design cycles were concluded to develop and optimize the most appropriate ML model in the context. **Result:** a ML model was developed using the SVM algorithm that obtained an AUC Precision-Recall equal to 0.660 and AUC ROC equal to 0.836. In addition, a minimum set of variables with higher predictive capacity than the complete set was identified. **Conclusion:** it was demonstrated, by using a supervised machine learning approach in conjunction with CGA, that it is possible to perform the risk classification of oncologic patients, in addition to optimize a SVM model to better identify high-risk patients.

**Keywords:** Machine Learning. Cancer. CGA.

# LISTA DE ILUSTRAÇÕES

Figura 1 – Idade do paciente ao ser diagnosticado com câncer. . . . .	20
Figura 2 – Realização da AGA no contexto do plano integrado de cuidado. . . . .	24
Figura 3 – Fluxo decisório após realização da AGA no IMIP. . . . .	26
Figura 4 – Exemplo de vazamento de dados durante o tratamento. . . . .	32
Figura 5 – Exemplo de matriz de confusão, usada na avaliação de algoritmos de classificação . . . . .	41
Figura 6 – Exemplo de curva ROC. . . . .	43
Figura 7 – Exemplo de curva Precision-Recall. . . . .	43
Figura 8 – Linha do tempo dos trabalhos relacionados. . . . .	46
Figura 9 – Ciclo de Engenharia. Os questionamentos indicam questões de conhecimento. As exclamações, problemas de design. . . . .	54
Figura 10 – Ciclo de design. Os questionamentos indicam questões de conhecimento. As exclamações, problemas de design. . . . .	55
Figura 11 – Exemplo de divisão dos dados durante o tratamento. . . . .	59
Figura 12 – Processo de desenvolvimento de modelos de AM para problemas de classificação. . . . .	60
Figura 13 – Etapas de pré-processamento definidas e executadas. . . . .	65
Figura 14 – Classificador de referência. . . . .	73
Figura 15 – Seleção de variáveis do início ao fim do processo. . . . .	77
Figura 16 – Curvas ROC e <i>Precision-Recall</i> para o algoritmo Máquina de Vetores de Suporte (SVM). . . . .	80
Figura 17 – Curvas ROC e <i>Precision-Recall</i> para o algoritmo Naive Bayes (NB). . .	82
Figura 18 – Curvas ROC e <i>Precision-Recall</i> para o algoritmo XGBoost. . . . .	82
Figura 19 – Curvas ROC e <i>Precision-Recall</i> para o algoritmo - Perceptron Multicamadas (MLP). . . . .	83
Figura 20 – Curvas ROC e <i>Precision-Recall</i> para o algoritmo Regressão Logística (LR). . . . .	83
Figura 21 – Curvas ROC e <i>Precision-Recall</i> para o algoritmo K Vizinhos Próximos (KNN). . . . .	84
Figura 22 – Curvas ROC, <i>Precision-Recall</i> e matriz de confusão (exemplo) para o algoritmo Análise Discriminante Linear (LDA). . . . .	85
Figura 23 – Curvas ROC, <i>Precision-Recall</i> para os modelos do ciclo 1 (cinza) e ciclo 2 (vermelho). . . . .	94
Figura 24 – Curvas ROC, Precision-Recall e matriz de confusão (exemplo) para o modelo desenvolvido utilizando o algoritmo SVM. . . . .	96

# LISTA DE TABELAS

Tabela 1 – Quantitativo de dados faltantes em variáveis categóricas.	67
Tabela 2 – Variáveis numéricas faltantes.	68
Tabela 3 – Variáveis categóricas codificadas removidas do dataset por apresentarem menos de 10 ocorrências.	69
Tabela 4 – Métricas obtidas no classificador de referência.	72
Tabela 5 – Resumo de desempenho dos algorítmicos avaliados no ciclo 1 de DSR.	80
Tabela 6 – Resumo de desempenho dos algoritmos avaliados no ciclo 2 de DSR. Após otimização realizada, houve melhora no desempenho em dados não vistos (dados de teste) em todos eles.	93

# LISTA DE QUADROS

Quadro 1 – Regra geral para interpretação dos coeficientes de correlação. . . . .	30
Quadro 2 – Exemplo de transformação de realizada em feature categórica (antes). . . . .	33
Quadro 3 – Exemplo de transformação de realizada em feature categórica (depois) . . . . .	33
Quadro 4 – Métricas de desempenho, derivadas da matriz de confusão. . . . .	41
Quadro 5 – Regra geral para interpretação dos valores da área sobre a curva ROC. . . . .	42
Quadro 6 – Quadro com tipos de artefatos. . . . .	55
Quadro 7 – Quadro com tipos, métodos e técnicas de avaliações de artefatos . . . . .	56
Quadro 8 – Quadro resumo com a quantidade de variáveis da AGA por cada categoria. . . . .	63
Quadro 9 – Variáveis removidas. . . . .	66
Quadro 10 – Itens identificados na etapa de investigação do problema (Ciclo 1) . . . . .	74
Quadro 11 – Seleção de variáveis. algoritmos BORUTA e RFE. . . . .	75
Quadro 12 – Resumo do número de variáveis utilizado por dimensão de avaliação. . . . .	77
Quadro 13 – Itens identificados na etapa de investigação do problema (Ciclo 2) . . . . .	86
Quadro 14 – Espaço de busca para o LDA. . . . .	87
Quadro 15 – Busca exaustiva realizada em 351 combinações diferentes de 3 parâmetros, além de validação cruzada. . . . .	87
Quadro 16 – Espaço de busca para a Regressão Logística. . . . .	88
Quadro 17 – Busca exaustiva realizada em 50 combinações diferentes de 2 parâmetros, além de validação cruzada. . . . .	88
Quadro 18 – Espaço de busca para o algoritmo Naive Bayes . . . . .	88
Quadro 19 – Busca exaustiva realizada em 9 combinações diferentes de 1 parâmetro, além de validação cruzada. . . . .	89
Quadro 20 – Espaço de busca para o algoritmo Máquina de Vetores de Suporte (SVM) . . . . .	89
Quadro 21 – Busca exaustiva realizada em 40 combinações diferentes de 2 parâmetros, além de validação cruzada. . . . .	90
Quadro 22 – Espaço de busca para o algoritmo KNN. . . . .	90
Quadro 23 – Busca exaustiva realizada em 72 combinações diferentes de 3 parâmetros, além de validação cruzada. . . . .	90
Quadro 24 – Espaço de busca para o algoritmo XGBoost. . . . .	91
Quadro 25 – Busca exaustiva realizada em 9360 combinações. . . . .	91
Quadro 26 – Espaço de busca para o algoritmo MLP. . . . .	92
Quadro 27 – Busca exaustiva realizada com 864 combinações de 5 parâmetros diferentes, além de validação cruzada. . . . .	92

Quadro 28 – Resumo do número de variáveis necessárias no algoritmo de AM por dimensão . . . . .	97
---	----

# LISTA DE ABREVIATURAS E SIGLAS

ADL/LDA	Análise Discriminante Linear / Linear Discriminant Analysis
AGA/CGA	Avaliação Geriátrica Ampla / Comprehensive Geriatric Assessment
AM	Aprendizagem de Máquina
DSR	Design Science Research
EORTC-QLQ30	European Organization for Research and Treatment of Cancer - Quality of Life Questionnaire
GDS	Geriatric Depression Scale
IMIP	Instituto de Medicina Integral Professor Fernando Figueira
KNN	K-nearest neighbors
MAN/MNA	Mini-Avaliação Nutricional / Mini-Nutritional Assessment
NB	Naive Bayes
PPS	Palliative Performance Scale
RFE	Recursive feature elimination
RL/LR	Rregressão Logística / Logistic Regression
SMOTE	Synthetic Minority Over-sampling Technique
SMOTE-NC	Synthetic Minority Over-sampling Technique-Nominal Continuous
SVM	Support Vector Machine
XGBOOST	eXtreme Gradient Boosting.

# SUMÁRIO

<b>1</b>	<b>INTRODUÇÃO</b>	<b>19</b>
<b>1.1</b>	<b>Contexto</b>	<b>19</b>
<b>1.2</b>	<b>Motivação</b>	<b>21</b>
<b>1.3</b>	<b>Problema</b>	<b>22</b>
<b>1.4</b>	<b>Objetivos</b>	<b>22</b>
<b>1.4.1</b>	Objetivo geral	22
<b>1.4.2</b>	Objetivos Específicos	22
<b>1.5</b>	<b>Estrutura do trabalho</b>	<b>22</b>
<b>2</b>	<b>REVISÃO DA LITERATURA</b>	<b>24</b>
<b>2.1</b>	<b>Avaliação Geriátrica Ampla (AGA)</b>	<b>24</b>
<b>2.2</b>	<b>Aprendizagem de máquina</b>	<b>27</b>
<b>2.2.1</b>	Pre-processamento dos dados	29
<b>2.2.2</b>	Seleção de variáveis ( <i>feature selection</i> )	29
<b>2.2.2.1</b>	Correlação	30
<b>2.2.2.2</b>	Remoção Recursiva de Variáveis ( <i>Recursive Feature Elimination - RFE</i> )	31
<b>2.2.2.3</b>	Algoritmo BORUTA	31
<b>2.2.3</b>	Vazamento de dados	31
<b>2.2.4</b>	Codificação de variáveis Categóricas	32
<b>2.2.5</b>	Normalização	34
<b>2.2.6</b>	Balanceamento entre classes	34
<b>2.2.7</b>	Algoritmos	35
<b>2.2.7.1</b>	Máquinas Vetoriais de Suporte (SVM)	35
<b>2.2.7.2</b>	Análise Discriminante Linear (LDA)	36
<b>2.2.7.3</b>	Naive Bayes	37
<b>2.2.7.4</b>	K-Vizinhos mais próximos	37
<b>2.2.7.5</b>	Regressão Logística	38
<b>2.2.7.6</b>	Rede Perceptron Multicamadas	38
<b>2.2.7.7</b>	Xgboost	39
<b>2.2.8</b>	Métricas de avaliação	40
<b>3</b>	<b>TRABALHOS RELACIONADOS</b>	<b>45</b>
<b>3.1</b>	<b>Visão Geral</b>	<b>45</b>
<b>3.2</b>	<b>Considerações Finais</b>	<b>50</b>
<b>4</b>	<b>METODOLOGIA</b>	<b>51</b>

4.0.1	Implementação utilizada . . . . .	51
<b>4.1</b>	<b>Fases da pesquisa . . . . .</b>	<b>51</b>
4.1.1	Revisão bibliográfica exploratória . . . . .	51
<b>4.2</b>	<b>Análise e tratamento dos dados . . . . .</b>	<b>51</b>
<b>4.3</b>	<b>Design Science Research (DSR) . . . . .</b>	<b>52</b>
4.3.1	Visão geral do DSR . . . . .	52
4.3.2	Ciclos do DSR . . . . .	52
4.3.3	Artefatos no DSR . . . . .	54
4.3.4	Métodos de avaliação de artefatos . . . . .	56
4.3.5	Aplicação do DSR nessa Pesquisa . . . . .	56
<b>5</b>	<b>ANÁLISE E TRATAMENTO DOS DADOS . . . . .</b>	<b>58</b>
<b>5.1</b>	<b>Decisões da pesquisa . . . . .</b>	<b>58</b>
5.1.1	Implementação utilizada . . . . .	58
5.1.2	Divisão do dataset . . . . .	58
5.1.2.1	Variável de saída: Definição da classe positiva . . . . .	59
5.1.3	Métricas utilizadas . . . . .	59
5.1.4	Processo de desenvolvimento de modelo de AM para classificação . . . . .	60
<b>5.2</b>	<b>Análise dos dados . . . . .</b>	<b>62</b>
5.2.1	Análise exploratória dos dados . . . . .	64
5.2.2	Pré-processamento dos dados . . . . .	64
5.2.2.1	Critérios de inclusão . . . . .	64
5.2.3	Remoção de variáveis sem informação ou com alta correlação . . . . .	65
5.2.3.1	Tratamento de dados faltantes . . . . .	67
5.2.3.2	Codificação das variáveis categóricas . . . . .	69
5.2.3.3	Remoção de variáveis categóricas raras . . . . .	69
5.2.3.4	Divisão dos dados entre treinamento/validação e teste . . . . .	71
5.2.3.5	Mudança de escala nas variáveis numéricas . . . . .	71
5.2.3.6	Balanceamento da base de dados . . . . .	71
<b>5.3</b>	<b>Modelo de Referência . . . . .</b>	<b>72</b>
<b>5.4</b>	<b>Conclusão . . . . .</b>	<b>73</b>
<b>6</b>	<b>CICLO 1 - SELEÇÃO DE FEATURES E DO ALGORITMO DE AM</b>	<b>74</b>
<b>6.1</b>	<b>Investigação do Problema . . . . .</b>	<b>74</b>
<b>6.2</b>	<b>Design da Solução . . . . .</b>	<b>74</b>
6.2.1	Seleção de variáveis ( <i>feature selection</i> ) . . . . .	75
6.2.2	Seleção de Algoritmos . . . . .	79
<b>6.3</b>	<b>Validação da Solução . . . . .</b>	<b>81</b>
<b>7</b>	<b>CICLO 2 - OTIMIZAÇÃO . . . . .</b>	<b>86</b>

<b>7.1</b>	<b>Investigação do Problema . . . . .</b>	<b>86</b>
<b>7.2</b>	<b>Design da Solução . . . . .</b>	<b>86</b>
7.2.1	Análise de Discriminante Linear . . . . .	87
7.2.2	Régressão Logística . . . . .	87
7.2.3	Naive Bayes . . . . .	88
7.2.4	Máquina de Vetores de Suporte . . . . .	89
7.2.5	K Vizinhos Próximos . . . . .	90
7.2.6	XGBoost . . . . .	91
7.2.7	Perceptron Multicamada . . . . .	91
<b>7.3</b>	<b>Validação da Solução . . . . .</b>	<b>92</b>
<b>8</b>	<b>CONCLUSÕES . . . . .</b>	<b>95</b>
<b>8.1</b>	<b>Considerações finais e contribuições . . . . .</b>	<b>95</b>
<b>8.2</b>	<b>Resultados alcançados . . . . .</b>	<b>98</b>
<b>8.3</b>	<b>Ameaças à validade . . . . .</b>	<b>98</b>
<b>8.4</b>	<b>Limitações da pesquisa . . . . .</b>	<b>98</b>
<b>8.5</b>	<b>Trabalhos futuros . . . . .</b>	<b>99</b>
	<b>REFERÊNCIAS . . . . .</b>	<b>100</b>
	<b>APÊNDICES</b>	<b>109</b>
	<b>APÊNDICE A – ARTIGO: MACHINE LEARNING APPLIED TO SURVIVAL PREDICTION OF ELDERLY CANCER PATIENTS: SYSTEMATIC REVIEW . . . . .</b>	<b>110</b>
	<b>APÊNDICE B – LISTAGEM COMPLETA DAS VARIÁVEIS DA AGA REALIZADA NO IMIP . . . . .</b>	<b>117</b>
	<b>ANEXOS</b>	<b>122</b>
	<b>ANEXO A – ANEXO: FORMULÁRIO IMIP COM AGA E QUESTIONÁRIO SOCIO-ECONÔMICO . . . . .</b>	<b>123</b>



# 1 INTRODUÇÃO

## 1.1 CONTEXTO

O termo *big data* surgiu em 1997 no artigo de Cox e Ellsworth que tratava do problema em lidar com dados que excediam a memória dos computadores disponíveis à época. Hoje, com os avanços na microeletrônica, o *big data* passou de problema à oportunidade das mais diversas. A definição atual, de acordo com a consultoria GARTNER (BEYER; LANEY, 2012), é que *big data* diz respeito a grandes volumes de dados, atualizados em alta velocidade e em grande variedade que necessitam de novos paradigmas de processamento para serem explorados. Ele não é caracterizado, por essa definição, por métricas específicas mas sim pelo fato de as abordagens tradicionais não serem capazes de processá-lo.

O uso de *big data* na medicina vem crescendo ano a ano e umas das áreas mais promissoras é a medicina de precisão ou medicina personalizada (CHIAVEGATTO, 2015), onde os testes diagnósticos, características biológicas, físicas, estilo de vida e até mesmo fatores genéricos são empregados para selecionar terapias apropriadas para cada paciente. (VERMA, 2012; YAU, 2019; COUNCIL, 2011).

De acordo com (CHIAVEGATTO, 2015), a maior parte do conhecimento médico é baseada em grandes médias, em escores de decisão que atendem a uma população. Por exemplo: um determinado tratamento reduz o risco de evento adverso em 80%. Significa que para 80% houve o efeito desejado (o risco foi eliminado), para os 20% restantes não houve o efeito esperado (não mitigou o risco). Atualmente, conhecemos o resultado para a população como um todo, mas não exatamente para quem. Com a medicina de precisão será possível identificar para quais pessoas esse tratamento surtirá, ou não, efeito. Como Chiavegatto (2015) exemplifica, pode ser que não funcione para mulheres acima de 60 anos, com pelo menos um filho, histórico de tabagismo, mutação em determinado gene e moradores de determinada região. A medicina de precisão fará com que, no futuro, ao invés de prescrever um medicamento para todos, prescrevê-lo apenas para os pacientes para os quais surtirá efeito.

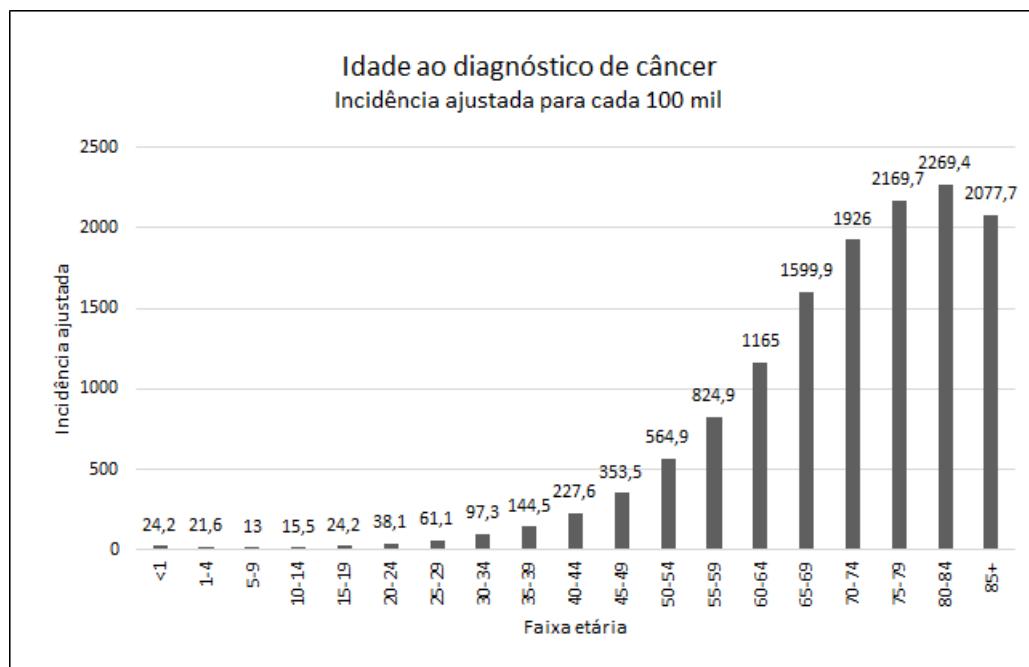
A aprendizagem de máquina (AM), derivada da Inteligência Artificial (IA), é a chave para a medicina de precisão pois lida com algoritmos capazes de aprender, encontrar padrões e realizar previsões com base em dados prévios (KOUROU et al., 2015). Ela tem sido utilizada ostensivamente na área médica, em conjunto com as grandes quantidades de dados sociais, psicológicos, genéticos, médicos, rotinas de tratamento, entre outros, que passaram a ser coletados e armazenados (MEDEIROS; LACERDA; BATISTA, 2019).

De acordo com a GLOBOCAN (BRAY et al., 2018), *Global Cancer Observatory*, o

câncer é um problema de saúde pública mundial, onde estima-se que, no ano 2030, haverão 27 milhões de casos incidentes, 17 milhões de mortes e 75 milhões de pessoas vivas, anualmente, com câncer. O maior efeito desse aumento vai incidir em países de baixa e média rendas.

De acordo com o relatório mais recente (abril/20, reportando o período de 1975 a 2017) do programa *Surveillance, Epidemiology, and End Results (SEER)* do *National Cancer Institute* americano (HOWLADER et al., 2020), a incidência de câncer ajustada por 100 mil americanos é de 223,4 para indivíduos abaixo de 65 anos e de 1956,7 para indivíduos acima de 65 anos. O gráfico da Figura 1 discrimina a incidência ajustada por faixa etária.

Figura 1 – Idade do paciente ao ser diagnosticado com câncer.



Fonte: Instituto Nacional do Câncer (EUA) *National Cancer Institute* (HOWLADER et al., 2020).

Portanto, câncer é uma doença do envelhecimento. Existem previsões que indicam que será a maior causa de morte em todo o mundo nos próximos anos e este número deve aumentar com o aumento da expectativa de vida da população (BRAY et al., 2018). A incidência de câncer está diretamente ligada à idade cronológica, porém somente a idade não é capaz de descrever a condição de saúde de uma pessoa.

Para completar este entendimento, utiliza-se a Avaliação Geriátrica Ampla (AGA) ou *Comprehensive Geriatric Assessment (CGA)* (WILDIERS et al., 2014; OWUSU; BERGER, 2014; EXTERMANN; HURRIA, 2007). Ela avalia os aspectos médicos, psicológicos e capacidades funcionais para desenvolver um plano integrado de tratamento, acompanhamento e reabilitação a longo prazo. A AGA é interdisciplinar pois considera não apenas os dados dos médicos, mas também dos enfermeiros e dos profissionais de saúde. É multidimensional pois considera não apenas os diagnósticos médicos, mas também as deficiências funcionais e

as questões ambientais e sociais que afetam o bem-estar dos pacientes. O tempo necessário para realizar todos os testes da AGA é de aproximadamente 45 minutos por paciente (WELSH, 2014; KENIS; WILDIERS, 2011; BGS, 2018).

A Sociedade Internacional de Oncologia Geriátrica, *International Society of Geriatric Oncology (SIOG)*, recomenda o uso da AGA para avaliação e acompanhamento de pacientes em tratamento contra o câncer, também destaca a existência de uma capacidade da AGA em prever desfechos clínicos, complicações ou mesmo óbito. (WILDIERS et al., 2014; OWUSU; BERGER, 2014; EXTERMANN; HURRIA, 2007).

## 1.2 MOTIVAÇÃO

A Avaliação Geriátrica Ampla (AGA) é o método mais adequado para obter uma visão sobre o estado geral de saúde de um paciente idoso, foi desenvolvido na medicina geriátrica como ferramenta diagnóstica, para planejar cuidados e intervenções. A AGA permite detectar múltiplos problemas que muitas vezes são desconhecidos para o oncologista de tratamento, permitindo também organizar intervenções específicas onde necessário (KENIS; WILDIERS, 2011).

Porém, conforme Ramjaun, Nassif e Krotneva (2013), a AGA não foi desenvolvida especificamente para o tratamento oncológico, por isso surgem algumas limitações que atrapalham a sua adoção: uma delas é o fato de envolver diversos profissionais de saúde. Outra é a duração de cerca de 45 minutos por paciente.

Por conta disto, uma estratégia em duas etapas normalmente é aplicada, onde na primeira é realizada uma triagem, com duração de poucos minutos. Caso seja detectado que o paciente é de risco, realiza-se a AGA para uma avaliação completa (EXTERMANN, 2010; MOHILE et al., 2007; OVERCASH et al., 2006).

Alguns centros utilizam formulários em papel, porém já existem versões eletrônicas da AGA como nos trabalhos de Sepehri et al. (2020) e Garm, Park e Song (2018). No entanto, apesar da realização da AGA, as decisões quanto à classificação de risco dos pacientes são realizadas com base nas grandes médias (ou escores), seja na versão eletrônica, seja na versão em papel.

De acordo com a revisão sistemática da literatura realizada (LACERDA et al., 2019), incluída no apêndice A, até então não existiam estudos publicados em língua inglesa utilizando a AGA em conjunto com modelos de AM para realizar previsões de sobrevida ou classificações de qualquer tipo.

## 1.3 PROBLEMA

Dante do exposto, o problema abordado neste trabalho está associado a seguinte pergunta de pesquisa:

**Como classificar os pacientes oncológicos em alto ou baixo risco de óbito em 6 meses utilizando as variáveis presentes na AGA?**

## 1.4 OBJETIVOS

Nesta Seção, detalham-se os objetivos geral e específicos desta pesquisa.

### 1.4.1 OBJETIVO GERAL

Desenvolver um modelo de aprendizagem de máquina capaz de classificar os pacientes idosos em tratamento oncológico em alto e baixo risco de morte em 6 meses, avaliando os dados disponíveis nos exames que compõem a Avaliação Geriátrica Ampla (AGA) e outras variáveis disponíveis.

### 1.4.2 OBJETIVOS ESPECÍFICOS

- Determinar qual o subconjunto de variáveis preditoras (*features*) capaz de realizar a classificação com a mesma qualidade, ou melhor, que o conjunto completo.
- Avaliar, de um conjunto de algoritmos de aprendizagem de máquina, o que apresenta o melhor desempenho.
- Otimizar a performance do algoritmo selecionado.
- Priorizar a identificação dos pacientes de alto risco.

## 1.5 ESTRUTURA DO TRABALHO

Além deste capítulo introdutório, o presente trabalho está organizado da seguinte forma:

- O capítulo 2 descreve a fundamentação teórica relacionada ao contexto do trabalho.
- O capítulo 3 apresenta e compara os trabalhos relacionados a esta pesquisa.
- O capítulo 4 destaca a metodologia utilizada.
- O capítulo 5 faz a análise e tratamento dos dados utilizados na pesquisa.

- Os capítulos 6 e 7 detalham a proposta por meio ciclos de *Design Science Research* (DSR).
- O capítulo 8 apresenta as conclusões e trabalhos futuros.
- O Apêndice A contém o artigo "Machine Learning Applied to survival prediction of elderly cancer patients: Systematic Review".
- O Anexo A contém o "Formulário IMIP com AGA e questionário Sócio-Econômico", que é aplicado aos pacientes da oncogeriatria do IMIP.

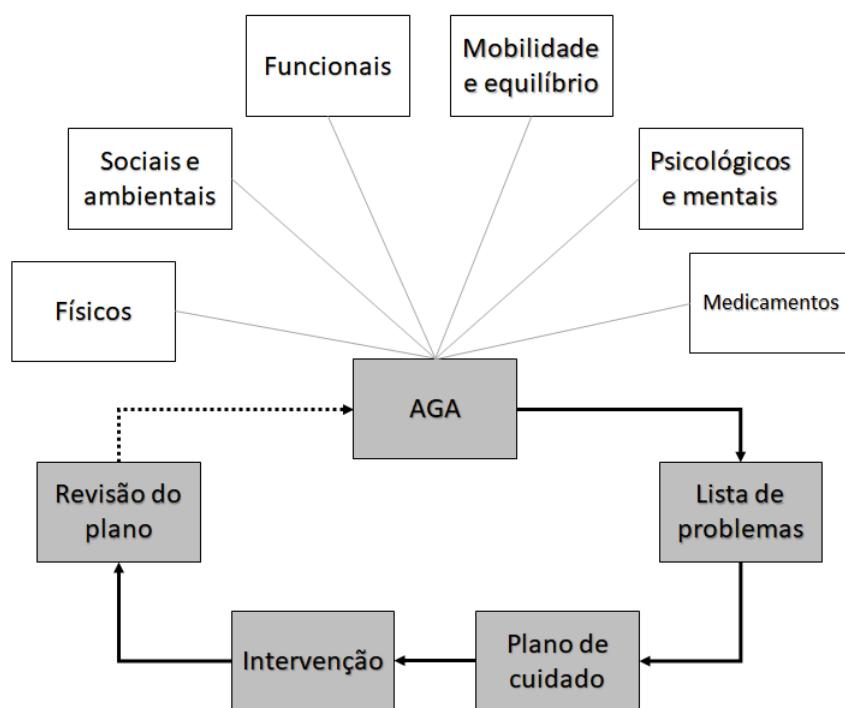
## 2 REVISÃO DA LITERATURA

Este capítulo, tem por objetivo apresentar os principais conceitos utilizados na nossa pesquisa, seja no campo da medicina, seja da Aprendizagem de Máquina (AM).

### 2.1 AVALIAÇÃO GERIÁTRICA AMPLA (AGA)

De acordo com a Sociedade Britânica de Geriatria, BGS - British Geriatrics Society (BGS, 2018), a AGA é definida como um processo diagnóstico multidimensional e interdisciplinar, que se baseia na determinação das capacidades médicas, psicossociais e funcionais de uma pessoa idosa para desenvolver um plano coordenado e integrado de tratamento e acompanhamento a longo prazo (Figura 2). Existem estudos (KENIS; WILDIERS, 2011; OVERCASH et al., 2019; WILDIERS et al., 2014) evidenciando a associação dos fatores avaliados na AGA com a previsão de sobrevida dos pacientes.

Figura 2 – Realização da AGA no contexto do plano integrado de cuidado.



Fonte: Sociedade Britânica de Geriatria (BGS, 2018).

No entanto, de acordo com Wildiers et al. 2014, os limiares para prever o agravamento ou modificar a abordagem terapêutica ainda não foram estabelecidos para diferentes tipos de câncer ou opções de tratamento.

Ainda, de acordo com a BGS 2018, os domínios avaliados pela AGA são (figura 2):

- **Físicos**

- Atributos físicos: peso, altura, IMC, sexo, idade e cor da pele.
- Doenças preexistentes: hipertensão arterial, AIDS, diabetes, hemiplegia, doença renal,
- Histórico de ocorrências médicas: infarto, insuficiência cardíaca, doença vascular cerebral, doença pulmonar crônica e doença do tecido conjuntivo.
- Estado nutricional, avaliado por meio do instrumento *Mini Nutritional Assessment* (GUIGOZ, 2006).
- Índice de comorbidades de Charlson (CHARLSON; POMPEI; ALES, 1987).

- **Sociais e ambientais**

- Histórico de tabagismo e consumo de álcool.
- Estado, cidade e bairro onde reside.
- Escolaridade, renda doméstica, estado civil, ocupação e religião.

- **Funcionais**

- Escala de performance paliativa (Paliative Performance Scale v2), (ANDERSON et al., 1996).
- Índice Karnofsky (KARNOFSKY, 1949).
- Índice KATZ-AVD - Avaliação de funcionalidade em atividades cotidianas (KATZ et al., 1970).
- Questionário QLQ-30 (*Quality of Life Questionnaire*). (FAYERS et al., 2001).

- **Mobilidade e equilíbrio**

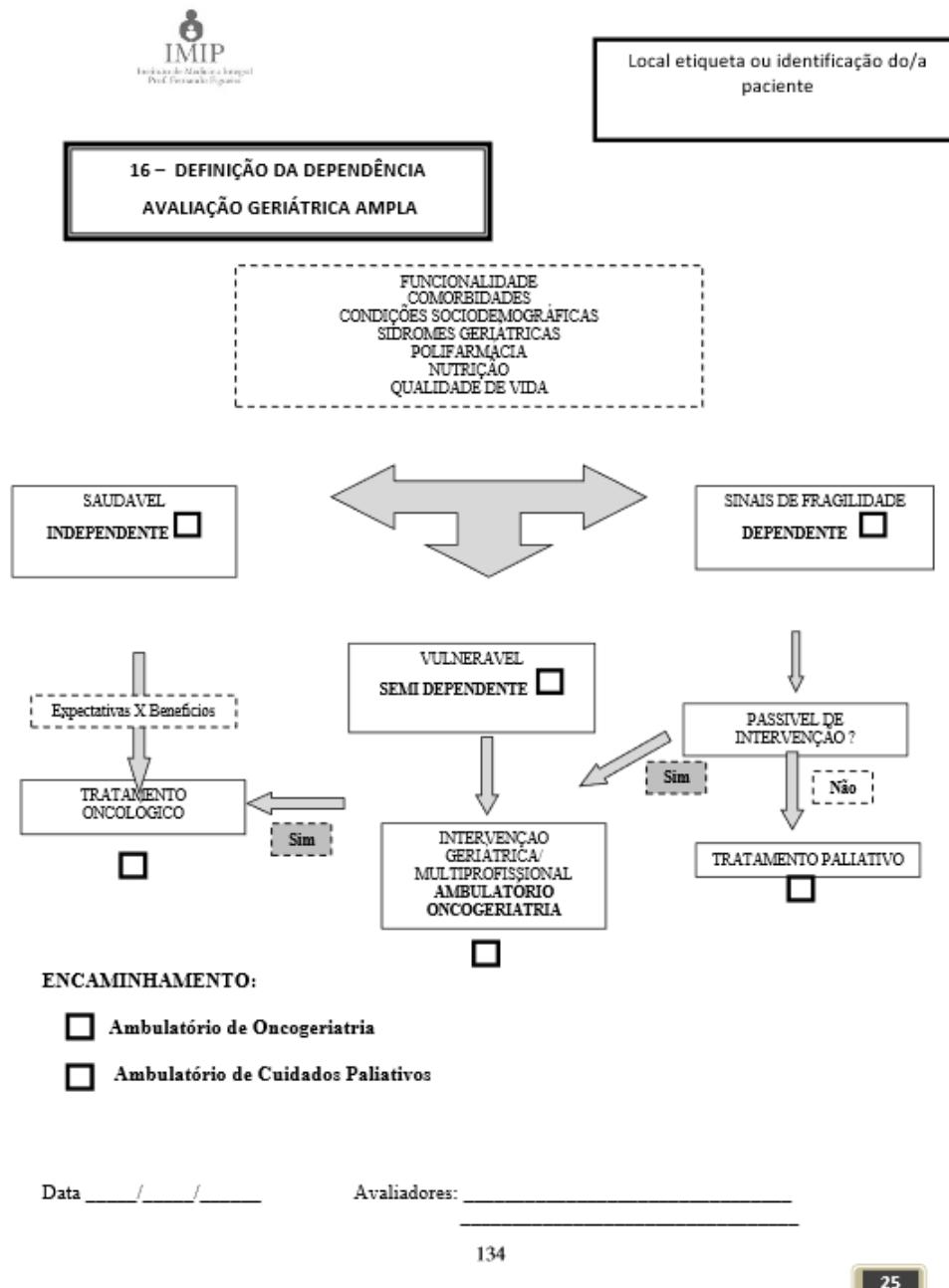
- Histórico de quedas.
- Teste *Timed up and go*. (MATHIAS; NAYAK; ISAACS, 1986).
- auto-avaliação física.

- **Psicológicos e mentais**

- Avaliação pela escala de depressão geriátrica (YESAVAGE; AL., 1982).
- Avaliação do estado cognitivo pelo teste mini mental (FOLSTEIN; FOLSTEIN; MCHUGH, 1975).

- **Medicamentos** Quantidade de medicamentos de uso continuo, receitados ou não, que o paciente utiliza.

Figura 3 – Fluxo decisório após realização da AGA no IMIP.



Fonte: AGA IMIP

O IMIP utiliza a versão da AGA que está presente no anexo A. De acordo com a avaliação, os pacientes são classificados em *saudáveis*, *vulneráveis* e *frágeis*. Os classificados como *saudáveis*, seguem para tratamento oncológico padrão. Os *vulneráveis*, seguem para tratamento oncológico porém com acompanhamento mais personalizado. Os *frágeis*, são reavaliados e toma-se a decisão se eles suportarão o tratamento oncológico ou se serão encaminhados para tratamento paliativo.

O mapa de decisão está na página 25, do anexo A, e reproduzido na Figura 3.

## 2.2 APRENDIZAGEM DE MÁQUINA

Por meio da Aprendizagem de Máquina (AM), os computadores são programados para induzir uma hipótese a partir de dados de experiências passadas. Isso só foi possível graças ao aumento da capacidade computacional, desenvolvimento de ferramentas (algoritmos de AM) e a disponibilidade de dados (*big data*). De acordo com Bishop (2006), os algoritmos de AM *aprendem* a induzir uma função ou hipótese capaz de resolver um problema a partir dos dados passados.

Segundo Bishop (2006), a AM é dividida em:

- **Aprendizado Supervisionado:** No aprendizado supervisionado o algoritmo experimenta os exemplos junto com os rótulos ou valores para cada exemplo. Os rótulos nos dados ajudam o algoritmo a correlacionar as variáveis de entrada. Duas das tarefas mais comuns de aprendizagem supervisionada de máquina são a **classificação** e a **regressão**, conforme descrito a seguir:
  - **Classificação:** Um problema de classificação é quando a variável de saída é categórica, como *vermelho* ou *preto* ou *paciente de risco alto* e *paciente de risco baixo*.
  - **Regressão:** Um problema de regressão é quando a variável de saída é um valor real, como por exemplo, um valor monetário.
- **Aprendizado não-supervisionado:** É quando não estão disponíveis os valores de saída para treinamento. O objetivo do aprendizado não-supervisionado é entender a estrutura ou distribuição dos dados. Eles são chamados não-supervisionados porque, ao contrário do aprendizado supervisionado, não existe resposta certa para aprender. Os algoritmos devem descobrir a estrutura dos dados.
- **Aprendizado semi-supervisionado:** Problemas onde dispomos de grandes quantidades de dados de entrada ( $X$ ) e apenas parte deles tem uma saída ( $Y$ ) mapeada. Um exemplo é um arquivo fotográfico onde algumas fotos estão rotuladas (por exemplo: gato ou cachorro) mas a maioria não está. Muitos problemas reais de AM são deste tipo. Isso ocorre porque normalmente é mais fácil conseguir grandes volumes de dados não rotulados do que rotulados.

No aprendizado supervisionado, o algoritmo de AM *aprende* um modelo por meio dos dados de treinamento. O objetivo do algoritmo é estimar a melhor função  $f$  para a variável de saída  $Y$ , dada a entrada  $X$ . Ainda segundo Bishop (2006), o erro da aproximação de qualquer algoritmo de AM pode ser dividido em duas partes:

- **Erro de Viés:** Viés são as suposições simplificadoras feitas por um modelo para facilitar a aprendizagem da função alvo. Geralmente os algoritmos mais simples têm um alto viés tornando-os rápidos de aprender e mais fáceis de entender, mas geralmente menos flexíveis. Por sua vez, eles têm menor desempenho preditivo em problemas complexos que não cumprem as suposições simplificadoras do viés dos algoritmos.
- **Erro de Variância:** É quanto que a estimativa da função alvo mudará se diferentes dados de treinamento forem utilizados. A função alvo é estimada a partir dos dados de treinamento por um algoritmo de AM, portanto devemos esperar que o algoritmo tenha alguma variância. Idealmente ele não deve mudar muito de um conjunto de dados de treinamento para o próximo, o que significa que o algoritmo é bom na escolha do mapeamento oculto subjacente entre as variáveis de entrada e saída. Algoritmos de aprendizado de máquina que possuem uma variância alta são fortemente influenciados pelas especificidades dos dados do treinamento, significa que os dados de treinamento têm influência sobre o número e tipos de parâmetros utilizados para caracterizar a função de mapeamento.

O objetivo do algoritmo de AM supervisionada é atingir baixo viés e baixa variância, porém há um *trade-off* entre viés e variância, conforme descrito por Bishop (2006), que é o conflito na tentativa de minimizar simultaneamente essas duas fontes de erro que impedem que algoritmos de aprendizagem supervisionada se generalizem além de seu conjunto de treinamento. Um alto viés acarreta um problema conhecido como *underfitting*, ao passo que a alta variância acarreta o *overfitting*, ambos descritos a seguir:

- **Overfitting:** Ocorre quando o algoritmo tem um bom desempenho apenas com dados de dados de treinamento, sendo pior com dados não vistos. Este fenômeno ocorre quando o modelo aprende os detalhes e ruídos dos dados de treinamento ao ponto de impactar negativamente o desempenho do modelo em novos dados, pois esses ruídos ou flutuações aleatórias são aprendidos como conceitos pelo modelo. O problema é que estes conceitos não se aplicam a novos dados e, por isso, impactam negativamente a capacidade do modelo de generalizar.
- **Underfitting:** Ocorre quando o modelo não foi capaz de aprender suficientemente o problema a partir dos dados de treinamento. Este modelo terá um mau desempenho já nos dados de treinamento, sendo de mais fácil identificação que o *overfitting*. Esse último apresentará bom desempenho com os dados de treinamento e um desempenho menor em dados não vistos.

De acordo com Bishop (2006), existem duas técnicas importantes que podem ser utilizadas ao avaliar algoritmos de AM e verificar se ocorre o *overfitting*, são elas:

- **Utilizar dados não vistos:** A melhor solução é avaliar a performance do modelo em dados que não foram utilizados durante o treinamento, porém nem sempre é possível dispor de novos dados. Uma alternativa é dividir a base de dados em treinamento e teste. Valores típicos são entre 67% a 80% dos dados para treinamento e os 33% a 20% restante para testes.
- **Validação cruzada (*k-fold cross validation*):** A técnica de validação cruzada consistem em dividir toda a base de dados em  $K$  partes, utilizar  $\frac{K-1}{K}$  partes para treinamento e a parte restante,  $\frac{1}{K}$ , para validação da performance. O processo é repetido  $K$ -vezes até que todas as  $K$ -partes tenham sido utilizadas como validação em algum momento. Valores típicos para  $K$  são entre 3 e 10.

### 2.2.1 PRE-PROCESSAMENTO DOS DADOS

Para o preprocessamento dos dados, descreve-se um modelo adaptado do proposto por Luo et al. (2016) e Chiavegatto et al. (2019), conforme descrito abaixo:

1. Definir os critérios de inclusão e exclusão de dados.
2. Remover variáveis sem informação ou com alta correlação.
3. Definir como será o tratamento de dados faltantes.
4. Codificar as variáveis categóricas.
5. Remover das variáveis categóricas raras (Identificar e remover variáveis independentes que apresentem predominantemente um único valor (ex. sendo zero 99% das amostras)).
6. Dividir da base de dados entre treinamento/validação e teste.
7. Mudar de escala nas variáveis numéricas.
8. Balancear a base de dados, com igual número de observações na variável dependente.

### 2.2.2 SELEÇÃO DE VARIÁVEIS (*FEATURE SELECTION*)

De acordo com Kohavi e John (1997), os algoritmos de AM apresentam uma performance inferior quando o número de variáveis utilizadas é maior que o ideal. Desta forma, deve-se buscar um conjunto de variáveis pequeno, idealmente mínimo, que garanta a melhor performance no problema em análise.

A seleção de variáveis é um processo onde seleciona-se automaticamente variáveis que mais contribuem para a variável de saída. Ter características irrelevantes nos dados poderá diminuir a precisão de muitos modelos, especialmente algoritmos lineares como regressão

linear e logística. De acordo com Brownlee (2019), três benefícios de realizar a seleção de variáveis antes de modelar seus dados são:

- **Reduz o sobreajuste (*overfitting*):** Dados menos redundantes significam menos oportunidade de tomar decisões baseadas em ruído.
- **Melhora a Precisão:** Menos dados enganosos significa que a precisão da modelagem melhora.
- **Reduz o tempo de treinamento:** Menos dados significam que os algoritmos treinam mais rápido.

Descreve-se abaixo três técnicas para seleção de variáveis: Por meio de coeficiente de Correlação, Remoção Recursiva de Variáveis (*Recursive Feature Elimination - RFE*) e o algoritmo conhecido como *BORUTA*.

#### 2.2.2.1 CORRELAÇÃO

Correlação é um método estatístico usado para avaliar a possível associação linear entre duas variáveis contínuas. Trata-se de uma relação de interdependência entre duas variáveis. É adimensional e varia de -1 (correlação negativa perfeita) passando por 0 (sem correlação) até +1 (correlação positiva perfeita). A correlação é medida por meio de coeficientes de correlação. Existem dois tipos principais de coeficientes de correlação: **Pearson e Spearman** (ALTMAN, 1990).

O coeficiente de correlação Pearson é utilizada quando ambas variáveis em estudo são normalmente distribuídas. Este coeficiente é afetado por valores extremos, dessa forma não é apropriada quando uma das variáveis não apresenta uma distribuição normal (ALTMAN, 1990).

De acordo com Hinkle, Wiersma e Jurs (2003), o coeficiente de correlação Spearman é mais robusto na presença de valores extremos ou quando uma ou ambas variáveis não apresentam distribuição normal. Abaixo, no Quadro 1, regra geral para interpretação dos coeficientes de correlação:

Quadro 1 – Regra geral para interpretação dos coeficientes de correlação.

Tamanho da correlação	Interpretação
.90 a 1.00 (-.90 a -1.00)	Correlação positiva (negativa) muito alta
.70 a .90 (-.70 a -.90)	Correlação positiva (negativa) alta
.50 a .70 (-.50 a -.70)	Correlação positiva (negativa) moderada
.30 a .50 (-.30 a -.50)	Correlação positiva (negativa) baixa
.00 a .30 (-.00 a -.30)	Correlação insignificante

Fonte: Mukaka (2012)

Manter duas variáveis com alta correlação no *dataset* de estudo reduz o desempenho durante o treinamento do modelo de AM, não contribui para melhorar a performance e, pelo contrário, introduz ruído que pode até, reduzir a performance (CHIAVEGATTO, 2019).

### 2.2.2.2 REMOÇÃO RECURSIVA DE VARIÁVEIS (*RECURSIVE FEATURE ELIMINATION - RFE*)

De acordo com Brownlee (2019), a Remoção Recursiva de Variáveis (RFE) remove recursivamente os atributos e constrói um modelo sobre aqueles atributos que permanecem. Utiliza a precisão do modelo para identificar quais atributos (e combinação de atributos) contribuem mais para a previsão da variável de saída. É necessário realizar uma busca exaustiva de 1 a N (N, sendo o total de variáveis da base de dados) para encontrar a subconjunto de features com melhor desempenho.

### 2.2.2.3 ALGORITMO BORUTA

Enquanto a abordagem anterior, RFE, busca encontrar o subconjunto mínimo de variáveis de entrada, o algoritmo desenvolvido por Rudnicki et al. (2006) e implementado por Kursa e Rudnicki (2010) é capaz de identificar todas as variáveis relevantes para o problema em questão. A determinação de todas as variáveis relevantes se faz importante para entender melhor o fenômeno em estudo (KURSA; RUDNICKI, 2010).

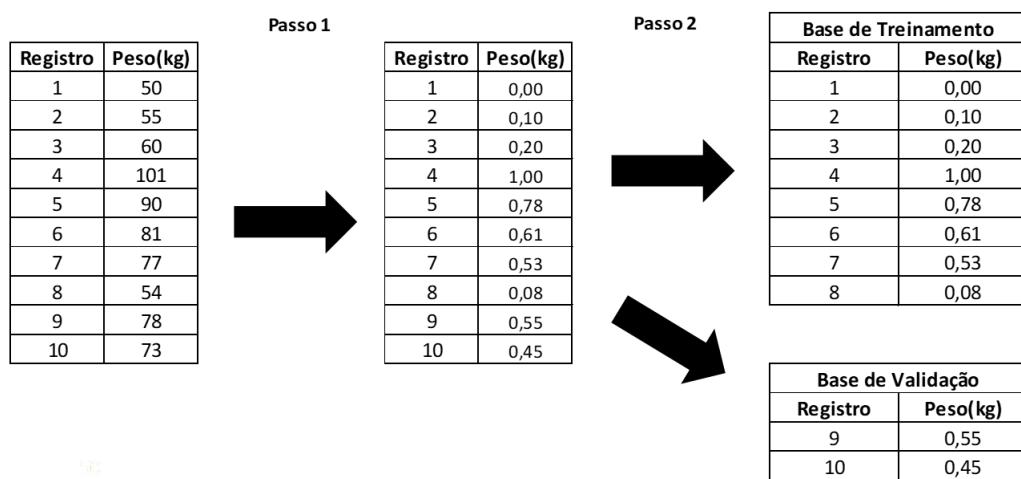
## 2.2.3 VAZAMENTO DE DADOS

De acordo com Kaufman, Rosset e Perllich (2011), o vazamento de dados (*data leakage*) é essencialmente a introdução de informação no modelo de AM que ele não deveria ter acesso, fazendo com que haja *overfitting* e, consequentemente, desempenho abaixo do esperado quando for apresentado à dados não vistos antes. Formalmente é: "A introdução não intencional de informação preditiva sobre a variável alvo durante o processamento, agregação e preparação dos dados (KAUFMAN; ROSSET; PERLICH, 2011)". Seguem dois exemplos abaixo:

**Exemplo 1:** Suponha que agregamos as bases de dados de pacientes de dois diferentes hospitais para criação de um modelo de AM. Porém, não desejamos manter na base a informação do hospital de origem, por isso removemos essa *feature*. Porém, se os registros dos pacientes do hospital A começam com "1" e os do hospital B começam com "2", a informação do hospital de origem terá vazado para o treinamento do modelo e poderá ser explorada pelo algoritmo de AM.

**Exemplo 2:** (Figura 4) Suponha uma base de dados com as features registro e peso(kg). A feature *peso* é submetida (passo 1) a um redimensionamento ( $X' = \frac{X - X_{min}}{X_{max} - X_{min}}$ ) e, posteriormente, é feita a divisão em base de treinamento e validação. O processo de redimensionamento de dados realizado teve conhecimento da distribuição completa dos dados ao calcular os fatores de escala (valor mínimo e máximo, neste caso. Poderia ser a média e desvio padrão). Esse conhecimento contaminou/vazou para os dados de validação, fazendo com que o algoritmo de AM implementado seja capaz de explorar essa vantagem indevida (KAUFMAN; ROSSET; PERLICH, 2011).

Figura 4 – Exemplo de vazamento de dados durante o tratamento.



Fonte: O autor

## 2.2.4 CODIFICAÇÃO DE VARIÁVEIS CATEGÓRICAS

Uma variável categórica é uma variável que apresenta uma quantidade limitada, normalmente fixa, de valores possíveis, atribuindo a cada observação a um dos valores possíveis (YATES; MOORE; STARNES, 2003).

Os algoritmos de AM esperam receber valores numéricos, portanto é necessário converter as variáveis categóricas em numéricas, de alguma forma. Os dois métodos mais simples de realizar esta conversão são descritos abaixo:

- **Codificação categórica simples (*Integer encoding*):** Nesta codificação, cada classe da variável categórica em questão é convertida para um valor numérico. Por exemplo, em uma variável categórica como *cor*, a codificação categórica simples vai atribuir um valor para cada classe como: Preto (0), Branco (1), Azul (2), Vermelho (3). Apesar dos algoritmos de AM já serem capazes de lidar com os dados após esta

conversão, o desempenho pode não ser o melhor, pois o algoritmo poderá dar mais peso a cor vermelha (por ter o maior valor).

- **Codificação categórica composta (*One-hot encoding*):** Este tipo de codificação cria novas variáveis binárias, uma para cada classe da variável categórica em questão.

Como exemplo, a variável categórica *ATV-Fisica-Classif* têm quatro classes diferentes a saber: [1] Muito ativo, [2] Ativo, [3] Irregularmente Ativo, [4] Sedentário, conforme Quadro 2. Ao realizar a codificação one-hot encoding cria-se 4 novas variáveis - cada uma para indicar a presença de um dos estados ou valores da variável conforme observa-se no Quadro 3.

Quadro 2 – Exemplo de transformação de realizada em feature categórica (antes).

Registro	ATV-Fisica-Classif
1	1
2	2
3	3
4	1
5	4
6	1
7	2

Fonte: O autor

Quadro 3 – Exemplo de transformação de realizada em feature categórica (depois)

Registro	ATV-Fisica-Cl_1	ATV-Fisica-Cl_2	ATV-Fisica-Cl_3	ATV-Fisica-Cl_4
1	1	0	0	0
2	0	1	0	0
3	0	0	1	0
4	1	0	0	0
5	0	0	0	1
6	1	0	0	0
7	0	1	0	0

Fonte: O autor

Variáveis com alta cardinalidade farão com que muitas novas features sejam criadas no dataset, no limite, pode ocorrer que a condição de que o número de variáveis preditoras (p) seja *muito menor* que a quantidade de amostras (n) ( $p \ll n$ ) não seja mais satisfeita. Isso acarreta que o volume do espaço p-dimensional cresce a ponto de tornar as n amostras disponíveis esparsas, problema conhecido como maldição da dimensionalidade (BELLMAN, 1961).

## 2.2.5 NORMALIZAÇÃO

Uma variável é normalizada através da mudança na escala dos seus valores de forma que estes fiquem dentro de uma pequena faixa especificada, tal como 0 a 1. A normalização pode melhorar a precisão e eficiência dos algoritmos AM, segundo Han e Kamber (2001), principalmente redes neurais, KNN e *clustering*. Para métodos baseados na distância, a normalização ajuda a evitar que atributos com gamas inicialmente grandes se sobreponham a atributos com gamas inicialmente menores (HAN; KAMBER, 2000).

Como exemplo: Imagine que na base de dados existam as variáveis preditoras peso e altura. O peso apresenta uma variação numérica (neste *dataset*) entre [50 e 150], já a altura varia entre [1 e 1.90]. Se não realizássemos a normalização, a variável peso iria se sobrepor a variável altura em importância.

- **Normalização mín-max:** Realiza a seguinte transformação: subtraí todo elemento do valor mínimo e divide pelo *span* (diferença entre o maior e o menor valor). Suponha que  $min_a$  e  $max_a$  são os valores mínimo e máximo para o atributo A. A normalização mín-max mapeia um valor X de A em X' no intervalo  $[min'_a, max'_a]$  por meio da transformação abaixo:

$$X' = \frac{X - min_a}{max_a - min_a}$$

Esse tipo de transformação preserva o aspecto original da distribuição, mas não reduz a importância dos valores extremos.

- **Normalização Escore padrão ou Z-score:** Os valores para um atributo A são normalizados com base na média e no desvio-padrão de A. Um valor X de A é normalizado para X' através do cálculo:

$$X' = \frac{X - \mu}{\sigma}$$

onde  $\mu$  e  $\sigma$  são respectivamente, a média e o desvio-padrão do atributo A. Este método de normalização é útil quando o mínimo e o máximo efetivos do atributo A são desconhecidos.

## 2.2.6 BALANCEAMENTO ENTRE CLASSES

Em aplicações práticas de problemas de classificação, normalmente há mais dados de uma determinada classe da variável de saída do que outra, fazendo com que algumas classes sejam subrepresentadas. A este fenômeno, da-se o nome de "desequilíbrio de classe", descrito por Prati, Batista e Monard (2009), que é o problema de aprender um conceito a partir da classe que tem um pequeno número de amostras. Segundo os trabalhos Yang e Wu.

(2006) e Rastgoo et al. (2016), o problema de desequilíbrio de classes tem sido encontrado em múltiplas áreas tais como telecomunicações, informática, bioinformática, detecção de fraudes e diagnóstico médico, tendo sido considerado um dos dez principais problemas para o uso de algoritmos de AM. Dados desbalanceados comprometem substancialmente o processo de aprendizagem, uma vez que a maioria dos algoritmos de AM espera uma distribuição de classes balanceada (HE; GARCIA., 2009).

Existem três abordagens para tratar o problema, são elas:

- **Subamostragem:** Refere-se ao processo de redução do número de amostras da classe que apresenta mais amostras.
- **Sobreamostragem:** Consiste em gerar novas amostras, seguindo um método, até se obter o equilíbrio entre as classes.
- **Combinação entre Sub e Sobreamostragem:** A sobreamostragem pode levar ao *overfitting*, enquanto a subamostragem pode levar a perda de informação. Essa abordagem apresenta-se como o equilíbrio entre as duas opções.

Os pesquisadores Lemaître, Nogueira e Aridas (2017), desenvolveram uma biblioteca em linguagem *python* que implementa o algoritmo para balanceamento entre classes, conhecido como *Synthetic minority over-sampling technique*, ou SMOTE. O SMOTE, por sua vez, foi desenvolvido por Chawla et al. (2002) bem como sua variante SMOTE-NC (*Nominal and Continuous features* - variáveis categóricas e numéricas), que é capaz de tratar variáveis categóricas e numéricas. Este algoritmo é capaz de implementar as três abordagens descritas acima para se equilibrar uma base de dados.

## 2.2.7 ALGORITMOS

Nesta seção, faremos uma breve descrição do algoritmos de AM testados neste trabalho.

### 2.2.7.1 MÁQUINAS VETORIAIS DE SUPORTE (SVM)

Segundo Smola e Schölkopf (2004), o SVM trabalha mapeando os dados para um espaço de características de alta dimensão para que os pontos de dados possam ser categorizados, mesmo quando os dados não são de outra forma linearmente separáveis. Um separador entre as categorias é encontrado, então os dados são transformados de tal forma que o separador poderia ser desenhado como um hiperplano.

Conforme analisado na literatura (BROWNLEE, 2019; BROWNLEE, 2016; SCIKIT-CLEAR, 2020), o Algoritmo de Máquinas Vetoriais de Suporte têm os seguinte hiperparâmetros de ajuste:

- **C:** Parâmetro de regularização. A força da regularização é inversamente proporcional à C. Deve ser estritamente positiva.
- **kernel:** Algoritmo utilizado na otimização. Quatro opções disponíveis: 'linear', 'poly', 'rbf', 'sigmoid'. O padrão é 'rbf'.

#### 2.2.7.2 ANÁLISE DISCRIMINANTE LINEAR (LDA)

*Linear Discriminant Analysis* ou LDA é uma técnica estatística para classificação binária e multiclasse. Ela também assume uma distribuição gaussiana para as variáveis numéricas de entrada. Segundo Brownlee (2016), este algoritmo é uma extensão da Regressão Logística, recomendando sempre testar ambos para saber qual melhor para o problema em estudo. O LDA também pode ser utilizado para realizar redução de dimensão da base de dados.

Conforme analisado na literatura (BROWNLEE, 2019; BROWNLEE, 2016; SCIKIT-CLEARN, 2020), o Algoritmo de Análise de Discriminante Linear (LDA) tem uma solução de forma fechada e, portanto, não tem hiperparâmetros para serem ajustados. Porém existem algumas opções que podem ser testadas.

De acordo com a documentação do Scikit-Clearn (2020), os parâmetros abaixo podem ser ajustados;

- **Solver:** É o algoritmo utilizado para otimizar a função. Três opções podem ser utilizadas
  - **svd:** *Singular value decomposition* (Valor padrão). Esse é o algoritmo recomendado para bases de dados com muitas features(centenas).
  - **lsqr:** Solução com mínimos quadrados.
  - **eigen:** Decomposição de autovalor.
- **Shrinkage:** ou retração, é uma ferramenta para melhorar a estimativa de matrizes de covariância em situações onde o número de amostras de treinamento é pequeno em comparação com o número de características. Não é o nosso caso. Assume os valores '*none*', '*auto*' ou qualquer valor entre 0 e 1.
- **Priors:** Informa ao algoritmo qual a proporção entre as classes, no modelo de classificação.
- **n\_components:** O LDA pode ser utilizado também para realizar a redução do número de dimensões da base de dados, porém como já realizamos este trabalho anteriormente, entendemos não ser necessário.

### 2.2.7.3 NAIVE BAYES

O algoritmo Naive Bayes calcula a probabilidade de cada classe e a probabilidade condicional de cada classe, dado cada valor de entrada. Estas probabilidades são estimadas para novos dados e multiplicadas em conjunto, assumindo que são todos independentes (uma suposição simples ou ingênuas).

Conforme analisado na literatura, (BISHOP, 2006; BROWNLEE, 2016; SCIKIT-CLEARN, 2020), o Algoritmo de Naive Bayes tem apenas um hiperparâmetro, **Priors**, que informa ao algoritmo qual a proporção entre as classes, no modelo de classificação.

### 2.2.7.4 K-VIZINHOS MAIS PRÓXIMOS

De acordo com Altman (1992), o algoritmo KNN encontra as distâncias entre uma consulta e todos os exemplos nos dados, selecionando os exemplos de números especificados ( $K$ ) mais próximos da consulta, depois vota para o rótulo mais frequente (no caso de classificação) ou faz a média dos rótulos (no caso de regressão). O número de amostras ( $K$ ) pode ser uma constante definida pelo usuário, ou variar de acordo com a densidade local dos pontos (aprendizagem do vizinho baseada no raio). A distância pode, em geral, ser qualquer medida métrica: a distância euclidiana padrão é a escolha mais comum.

Apesar de sua simplicidade, os vizinhos mais próximos têm tido sucesso em um grande número de problemas de classificação e regressão, incluindo dígitos escritos à mão e cenas de imagem de satélite. Sendo um método não paramétrico, muitas vezes é bem sucedido em situações de classificação onde o limite de decisão é muito irregular.

Conforme analisado na literatura, (BROWNLEE, 2019; BROWNLEE, 2016; SCIKIT-CLEARN, 2020), o algoritmo KNN têm os seguintes hiperparâmetros de ajuste:

- **n\_neighbors:** Número de vizinhos para ser considerado. Valor padrão é igual 5.
- **weights:** Aceita os valores abaixo:
  - **uniform:** Pesos uniforme para os vizinhos.
  - **distance:** Atribui pesos aos vizinhos de forma inversamente proporcional a distância, ou seja, os vizinhos próximos tem maior influência.
- **algorithm:** Possui 3 alternativas(além da forma *auto*):
  - **ball\_tree:** Utiliza o algoritmo Ball Tree.
  - **kd\_tree:** Utiliza o algoritmo KD Tree.
  - **brute:** Utiliza uma abordagem de força Bruta.

### 2.2.7.5 REGRESSÃO LOGÍSTICA

Segundo Brownlee (2016), a regressão logística, apesar de seu nome, é um modelo linear de classificação e não de regressão. A regressão logística também é conhecida na literatura como regressão logit. Neste modelo, as probabilidades descrevendo os possíveis resultados de um único estudo são modeladas através de uma função logística.

É um algoritmo de classificação, que é utilizado onde a variável resposta é categórica. A ideia da Regressão Logística é encontrar uma relação entre características e probabilidade de um determinado resultado.

Conforme analisado na literatura, (BROWNLEE, 2019; BROWNLEE, 2016; SCIKIT-CLEARN, 2020), o algoritmo de regressão logística têm os seguintes hiperparâmetros de ajuste:

- **penalty:** A ideia básica da penalização é evitar o overfitting impondo um *amortecimento* ou *penalização* para mudanças bruscas nos parâmetros. Existem 4 algoritmos disponíveis ‘l1’, ‘l2’, ‘elasticnet’, ‘none’.
- **C:** Parâmetro que define a força da regularização. Valor padrão é igual a 1.
- **class\_weight:** Informa ao algoritmo qual a proporção entre as classes, no modelo de classificação.
- **solver:** Algoritmo utilizado na otimização. Quatro opções disponíveis: ‘newton-cg’, ‘lbfgs’, ‘liblinear’, ‘sag’ e ‘saga’. O valor padrão é ‘lbfgs’.

### 2.2.7.6 REDE PERCEPTRON MULTICAMADAS

Segundo Bishop (2006), o campo de redes neurais artificiais é muitas vezes chamado apenas de Redes Neurais ou Perceptron Multicamadas. A base é o perceptron que é um modelo de neurônio único, desenvolvido por Rosenblatt (1958), e que foi um precursor de redes neurais maiores. É um campo de estudo que investiga como modelos simples de cérebros biológicos podem ser usados para resolver tarefas computacionais difíceis como as tarefas de modelagem preditiva que vemos na aprendizagem de máquinas. O objetivo não é criar modelos realistas do cérebro, mas sim desenvolver algoritmos robustos e estruturas de dados que possamos usar para modelar problemas difíceis. O poder das redes neurais vem de sua capacidade de aprender a representação em seu dados de treinamento e como melhor relacioná-los com a variável de saída desejada. Neste sentido, as redes neurais aprendem um mapeamento. Matematicamente, elas são capazes de aprender qualquer função de mapeamento e têm demonstrado ser um algoritmo universal de aproximação. A capacidade preditiva das redes neurais vem da estrutura hierárquica ou multicamadas das redes.

Um campo de estudo é como definir a arquitetura da rede neural para resolver um determinado problema. Segundo Walczak e Cerpa (1999), redes muito complexas tentam ao sobreajuste (*overfitting*) enquanto que arquiteturas mais simples não contêm elementos de processamento suficientes para modelar corretamente o conjunto de dados de entrada, ocorrendo então o subajuste (*underfitting*). Ambas as situações resultam em baixa performance. Walczak e Cerpa (1999), definiram um conjunto de heurísticas para guiar design de redes neurais artificiais.

Conforme analisado na literatura por (BROWNLEE, 2019; BROWNLEE, 2016; SCIKIT-CEARN, 2020), o algoritmo Perceptron Multicamadas têm os seguintes hiperparâmetros de ajuste:

- **hidden\_layer\_sizes:** tupla representando a quantidade de neurônios por camada oculta da rede neural. Definir o tamanho e quantidade de neurônios por camada impacta diretamente na performance do modelo. Como dito anteriormente, existem heurísticas para guiar o processo de design da rede neural.
- **activation:** Tipo de função de ativação para os neurônios das camadas ocultas. Admite os valores:
  - **identity:** Sem função de ativação.
  - **logistic:** Função logística(sigmoide).
  - **tanh:** Função tangente hiperbólica.
  - **relu:** Função retificada linear.
- **solver:** Algoritmo utilizado na otimização. Três opções disponíveis: ‘lbfgs’, ‘sgd’, ‘adam’. O padrão é ‘adam’.
- **alpha:** Parâmetro de *amortização* ou *penalidade*.
- **learning\_rate:** Taxa de atualização dos pesos. Assume três valores: ‘constant’, ‘invscaling’ ou ‘adaptive’.

#### 2.2.7.7 XGBOOST

O XGBoost, Algoritmo desenvolvido por Chen e Guestrin (2016), é a abreviação de *Extreme Gradient Boosting* e é uma implementação eficiente do algoritmo de aprendizagem de máquinas de aumento de gradiente estocástico. É um conjunto de algoritmos de árvores de decisão onde novas árvores corrigem os erros das árvores que já fazem parte do modelo onde árvores são adicionadas até que nenhuma outra melhoria possa ser feita no modelo. O XGBoost fornece uma implementação altamente eficiente do algoritmo de aumento do gradiente estocástico.

Conforme analisado na literatura (BROWNLEE, 2019) e a documentação do XGBoost (XGBOOST, 2020), o algoritmo têm os seguintes hiperparâmetros de ajuste:

- **scale\_pos\_weight:** Informa ao algoritmo qual a proporção entre as classes, no modelo de classificação.
- **learning\_rate:** Taxa utilizada para atualizar os pesos do algoritmo. Quanto menor, mais lenta é a taxa de atualização. Valor padrão é igual a 0.3
- **max\_depth:** Profundidade máxima da árvore de busca. Aumentar este valor fará o modelo ficar mais complexo e mais propenso ao *overfitting*. Valor padrão é igual a 6.
- **subsample:** Parâmetro utilizado para descartar algumas amostras. Utilizado para prevenir *overfitting*.
- **colsample\_bytree:** É a relação de subamostragem das colunas na construção de cada árvore. A subamostragem ocorre uma vez para cada árvore construída.
- **gamma:** Redução mínima das perdas necessárias para fazer uma nova partição em um nó de folha da árvore. Quanto maior for a gama, mais conservador será o algoritmo.

## 2.2.8 MÉTRICAS DE AVALIAÇÃO

Abaixo conceitos e medidas mais utilizadas de acordo com Kumari e Kishore (2018) e resumidas no Quadro 4.

- **Verdadeiro Positivo (VP):** Representa os casos positivos, que foram corretamente previstos como positivos.
- **Verdadeiro Negativo (VN):** Representa os casos negativos, que foram corretamente previstos como negativos.
- **Falso Positivo (FP):** Representa os casos negativos, que foram incorretamente previstos como positivos.
- **Falso Negativo (FN):** Representa os casos positivos, que foram incorretamente previstos como negativos.
- **Acurácia:** Proporção de observações classificadas corretamente. Não é muito útil em problemas com classes desbalanceadas.
- **Matriz de confusão:** É a representação dos valores acima: VP, VN, FP e FN em uma tabela conforme Figura 5.

Quadro 4 – Métricas de desempenho, derivadas da matriz de confusão.

Métrica	Fórmula Matemática	Descrição
Negativos	$VN + FP$	Número de amostras Negativas
Positivos	$VP + FN$	Número de amostras Positivas
Acurácia	$\frac{VP+VN}{VP+VN+FP+FN}$	Proporção de observações classificadas corretamente
Precisão	$\frac{VP}{VP+FP}$	Proporção de observações positivas classificadas corretamente como verdadeiras
Recall TVP	$\frac{VP}{VP+FN}$	Proporção de classificações da classe positiva que estão corretas
TFP	$\frac{FP}{VN+FP}$	Proporção de classificações da classe negativa que estão corretas
Especificidade	$\frac{VN}{FP+VN}$	Proporção de verdadeiros negativos corretamente identificados.
F1-Score	$2 \cdot \left( \frac{precisao \cdot recall}{precisao + recall} \right)$	Média harmônica entre precisão e recall

Fonte: O Autor.

Figura 5 – Exemplo de matriz de confusão, usada na avaliação de algoritmos de classificação

		n		p
		Resposta Preditiva		
n	n	VN	FP	
	p	FN	VP	

Fonte: O Autor.

- **Precisão:** Número de verdadeiros positivos (VP) dividido pelo total de verdadeiros (VP) e falsos positivos (FP). Não são considerados os Falso Negativos (FN). Utilizada quando é apropriado minimizar os falso positivos (FP).
- **Taxa de Verdadeiros Negativos ou Especificidade:** É Proporção de verdadeiros negativos corretamente identificados. Sumariza o quanto bem a classe negativa foi prevista.
- **Taxa de Verdadeiros Positivos (TVP), Sensibilidade, Revocação ou Recall:** Recall é uma métrica que quantifica o número de previsões positivas corretas feitas a partir de todas as previsões positivas corretas que poderiam ter sido feitas. Ao contrário da precisão que apenas comenta as previsões positivas corretas de

todas as previsões positivas, o recall fornece uma indicação das previsões positivas perdidas. Desta forma, o recall fornece alguma noção da cobertura da classe positiva. Matematicamente, é o número de verdadeiros positivos (VP) dividido pela soma de verdadeiros positivos (VP) e falsos negativos (FN). Utilizada quando é apropriado minimizar os falso negativos (FN).

- **Taxa de Falsos Positivos (TFP):** Matematicamente, é o número de falsos positivos (FP) dividido pela soma de verdadeiros positivos (VP) e falsos negativos (FN).
- **F1-Score:** Média harmônica entre entre acurácia e a sensibilidade. Proporciona uma forma de combinar a precisão e o *recall* em um único valor.

As curvas ROC e *precision-recall* provém um diagnóstico para os modelos de classificação binária onde a área sob essas curvas são utilizadas como escores que as summarizam e podem ser utilizados para comparar a performance de classificadores.

- **Área sobre curva ROC - AUC ROC:** Uma eficiente medida de desempenho de modelos de AM é a Área sobre a Curva ROC. A curva ROC, de *Receiver Operating Characteristic*, é obtida a partir da matriz de confusão, onde o eixo horizontal indica a taxa de falsos positivos (1-especificidade) e o eixo vertical indica a taxa de verdadeiros positivos (*recall*). Cada ponto no espaço representa os respectivos valores obtidos de uma matriz de confusão. A AUC ROC varia de 0 (zero) a 1 (um), com 1 indicando um classificador perfeito. Porém, mesmo amplamente utilizada, a AUC ROC tem suas limitações. De acordo com Fernández et al. (2018), para problemas com classes desbalanceadas, com poucos exemplos na classe minoritária, as estimativas podem não ser confiáveis (Figura 6).

No Quadro 5 define-se uma regra geral para interpretação dos valores da ROC AUC, conforme Kumari e Kishore (2018):

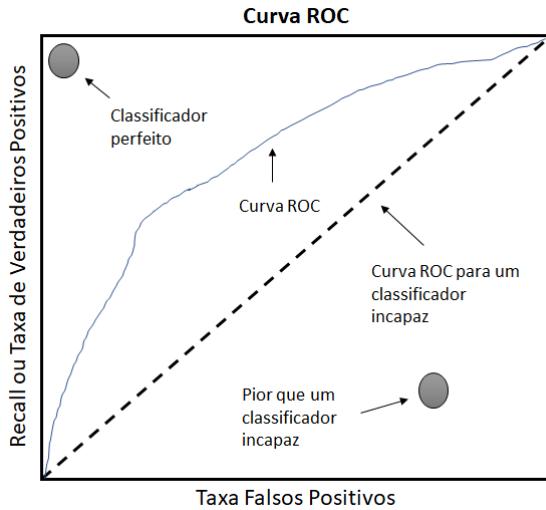
Quadro 5 – Regra geral para interpretação dos valores da área sobre a curva ROC.

Valor AUC	Interpretação
0.9 a 1	Discriminação excelente
0.8 a 0.9	Discriminação boa
0.7 a 0.8	Discriminação aceitável
0.6 a 0.7	Discriminação ruim

Fonte: Kumari e Kishore (2018).

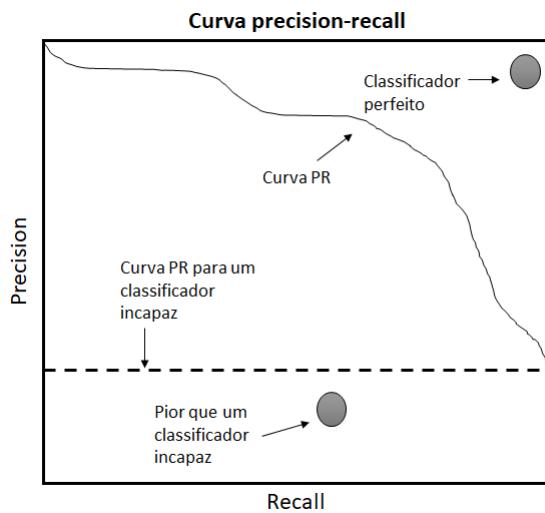
- **Área sobre a curva precisão-revocação ou Precision-Recall:** Uma alternativa à AUC ROC e a AUC *precision-recall*, pois ele indica mais precisamente a

Figura 6 – Exemplo de curva ROC.



Fonte: O autor

Figura 7 – Exemplo de curva Precision-Recall.



Fonte: O Autor.

performance da classificação na classe positiva (classe minoritária) pois tanto a precisão quanto o *recall* são métricas com ênfase na classe positiva (Figura 7).

A *precisão* pode ser usada em uma situação em que os Falsos Positivos são considerados mais prejudiciais que os Falsos Negativos. Por exemplo, ao classificar uma ação como um bom investimento, é necessário que o modelo esteja correto, mesmo que acabe classificando bons investimentos como maus investimentos (situação de Falso Negativo) no processo. Ou seja, o modelo deve ser preciso em suas classificações, pois a partir do momento que consideramos um investimento bom quando na verdade ele não é, uma grande perda de dinheiro pode acontecer.

O *recall* pode ser usada em uma situação em que os Falsos Negativos são considerados mais prejudiciais que os Falsos Positivos. Por exemplo, o modelo deve de qualquer maneira encontrar todos os pacientes doentes, mesmo que classifique alguns saudáveis como doentes (situação de Falso Positivo) no processo. Ou seja, o modelo deve ter alto recall, pois classificar pacientes doentes como saudáveis é prejudicial.

### 3 TRABALHOS RELACIONADOS

Este capítulo apresenta os trabalhos relacionados com esta pesquisa, explicitando os critérios que foram usados para esta seleção e suas fontes de pesquisa. Ao fim deste capítulo é apresentada uma discussão sobre os trabalhos selecionados, juntamente com uma comparação destes sob os critérios relevantes para esta pesquisa.

#### 3.1 VISÃO GERAL

Realizou-se uma revisão sistemática compreendendo o período entre janeiro de 2008 e agosto de 2019 e que encontra-se disponível no anexo A sob forma de artigo científico publicado na *15<sup>a</sup> Conferencia Ibérica de Sistemas y Tecnologías de Información - CISTI '2020*, na categoria artigo (*full paper*).

A questão principal da pesquisa realizada na revisão sistemática foi:

- Como a aprendizagem de máquina pode ser aplicada à predição de sobrevida de pacientes em tratamento contra o câncer?

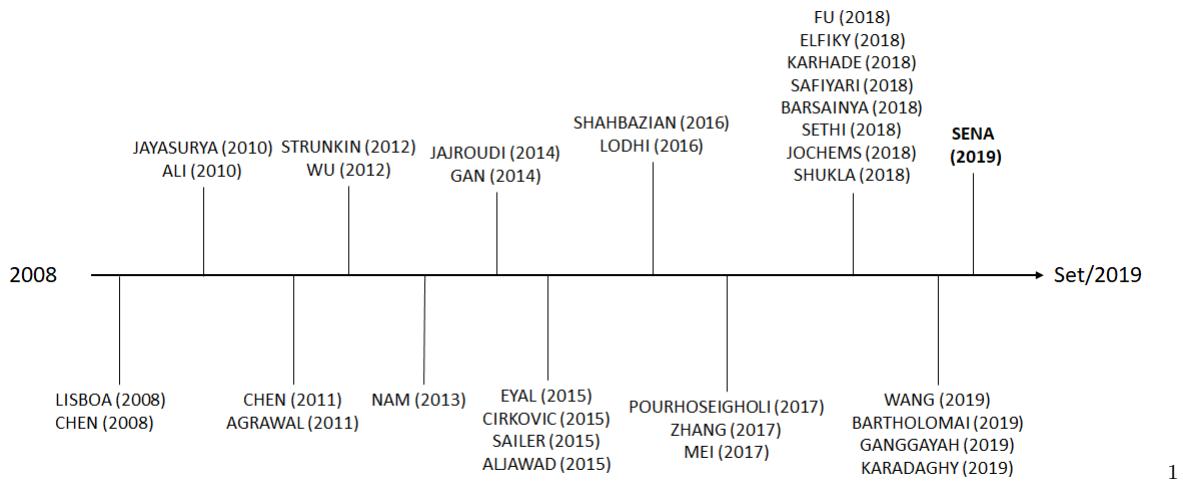
As perguntas secundárias, utilizadas para melhorar o entendimento, e guiar a revisão, foram:

- Quais variáveis podem ser utilizadas na predição?
- É possível utilizar as informações da AGA/CGA para realizar predições?
- Como categorizar os pacientes com relação ao risco de morte?
- Quais algoritmos e técnicas de Aprendizagem de Máquina utilizadas?
- Quais as oportunidades e desafios relacionados ao uso de AM e a predição de sobrevida em pacientes em tratamento contra o câncer?

A pesquisa foi realizada nos cinco portais científicos descritos abaixo:

- ACM (Association for Computing Machinery) Digital Library
- IEEE Xplore
- ScienceDirect – Elsevier
- PubMed

Figura 8 – Linha do tempo dos trabalhos relacionados.



Fonte: O autor.

- Journals of Clinical Oncology

A *string* de busca utilizada foi:

- ("machine learning"OR "data mining"OR "data science") AND ("cancer") AND ("prediction")

Encontrou-se, nestas bases, 1855 artigos únicos que, após crivo descrito no nosso artigo (constante no anexo A), restaram 32 artigos para análise, descritos a seguir e representados na linha do tempo da Figura 8.

- “Time-to-event analysis with artificial neural networks: An integrated analytical and rule-based study for breast cancer”, (LISBOA et al., 2008).
- “A clustering approach in developing prognostic systems of cancer patients”, (CHEN et al., 2008).
- “Analyzing potential of SVM based classifiers for intelligent and less invasive breast cancer prognosis”, (ALI et al., 2010).
- “Comparison of Bayesian network and support vector machine models for two-year survival prediction in lung cancer patients treated with radiotherapy”, (JAYASURYA et al., 2010).
- “Prediction of survival in patients with liver cancer using artificial neural network and classification and regression trees”, (CHEN et al., 2011).

- “An investigation into feature selection for oncological survival prediction”, (STRUNKIN; NAMEE; KELLEHER, 2012).
- “An examination of TNM staging of melanoma by a machine learning algorithm”, (WU et al., 2012).
- "A Hybrid Cancer Prognosis System Based on Semi-Supervised Learning and Decision Trees", (NAM; SHIN, 2013).
- “A lung cancer outcome calculator using ensemble data mining on SEER data”, (AGRAWAL et al., 2011)
- “A survey of pattern classification based methods for predicting survival time of lung cancer patients”, (GAN; ZHENG; WANG, 2014).
- “Prediction of survival in thyroid cancer using data mining technique”, (JAJROUDI et al., 2014).
- “Prediction of 5-year survival with data mining algorithms”, (SAILER et al., 2015).
- ”Comparison of three classifiers for breast cancer outcome prediction”, (EYAL; LAST; RUBIN, 2015).
- “Prediction models for estimation of survival rate and relapse for breast cancer patients”, (CIRKOVIC et al., 2015).
- “A framework to predict outcome for cancer patients using data from a nursing EHR”, (LODHI et al., 2016).
- “Predictive model for survival in patients with gastric cancer”, (SHAHBAZIAN; JAFARI; HAGHNIA, 2016).
- “Breast cancer surgery survivability prediction using Bayesian network and support vector machines”, (ALJAWAD et al., 2017).
- “Comparison of basic and ensemble data mining methods in predicting 5-year survival of colorectal cancer patients”, (POURHOSEINGHOLI; KHEIRIAN; ZALI, 2017).
- “Using the machine learning approach to predict patient survival from high-dimensional survival data”, (ZHANG; TANG; WANG, 2017).
- “Predicting five-year overall survival in patients with nonsmall cell lung cancer by reliefF algorithm and random forests”, (MEI, 2017).
- “Predicting lung cancer survivability using ensemble learning methods”, (SAFIYARI; JAVIDAN, 2018).

- “Development of Machine Learning Algorithms for Prediction of 5-Year Spinal Chordoma Survival”, (KARHADE et al., 2018).
- “Analysis and Prediction of Survival after Colorectal Chemotherapy using Machine Learning Models”, (BARSAINYA; SAIRAM; PATIL, 2018).
- “A prediction model for early death in non-small cell lung cancer patients following curative intent chemoradiotherapy”, (JOCHEMS et al., 2018).
- “Breast cancer data analysis for survivability studies and prediction”, (SHUKLA et al., 2018a).
- ”Applying machine learning in cancer prognosis using expression profiles of candidate genes”, (FU; CHENG; DING, 2018).
- “Analogizing of Evolutionary and Machine Learning Algorithms for Prognosis of Breast Cancer”, (SETHI, 2018).
- “Development and Application of a Machine Learning Approach to Assess Short term Mortality Risk Among Patients With Cancer Starting Chemotherapy”, (ELFIKY et al., 2018).
- “Lung Cancer Survival Prediction via Machine Learning Regression, Classification, and Statistical Techniques”, (BARTHOLOMAI; FRIEBOES, 2019).
- “Predicting factors for survival of breast cancer patients using machine learning techniques”, (GANGGAYAH et al., 2019).
- “Development and Assessment of a Machine Learning Model to Help Predict Survival among Patients with Oral Squamous Cell Carcinoma”, (KARADAGHY et al., 2019).
- “A tree ensemble based two-stage model for advanced stage colorectal cancer survival prediction”, (WANG et al., 2019).

Além destes trabalhos, selecionamos o seguinte artigo, publicado após o período da revisão sistemática:

- "Developing Machine Learning Algorithms for the Prediction of Early Death in Elderly Cancer Patients: Usability Study", (SENA et al., 2019).

Na nossa revisão, analisam-se os artigos selecionados nas dimensões abaixo discriminadas.

- Tipo de câncer.

- Total de pacientes.
- Algoritmos testados.
- Tipo de predição realizada.
- Desempenho obtido.
- Variáveis utilizadas, mas especificamente se utilizou a AGA.

Neste capítulo, analisou-se quais as variáveis utilizadas para o desenvolvimento do modelo de AM, dimensão considerada a mais relevante. Neste sentido, observa-se uma pluralidade de tipos de variáveis utilizadas, conforme descrito a seguir.

Dos 33 trabalhos analisados, apenas Sena et al. (2019) utilizou a AGA em conjunto com dados sociodemográficos para realizar previsões de óbito em 6 meses. Lodhi et al. (2016), utilizou um banco de dados multi-cêntrico com informações geridas pela equipe de enfermagem, que mais se assemelha a AGA.

Mei (2017), Cirkovic et al. (2015) e Chen et al. (2008) utilizaram dados clínicos e histopatológicos para realizar previsões.

Os 15 trabalhos a seguir utilizaram dados demográficos dos pacientes em conjunto com dados clínicos e histopatológicos: (ALI et al., 2010), (AGRAWAL et al., 2011), (BARTHOLOMAI; FRIEBOES, 2019), (CHEN et al., 2011), (NAM; SHIN, 2013), (JAJROUDI et al., 2014), (SHAHBAZIAN; JAFARI; HAGHNIA, 2016), (BARSAINYA; SAIRAM; PATIL, 2018), (JOCHEMS et al., 2018), (ELFIKY et al., 2018), (SHUKLA et al., 2018b), (WANG et al., 2019), (KARADAGHY et al., 2019) e (GANGGAYAH et al., 2019). Destaque para (ELFIKY et al., 2018), que adicionou uma dimensão temporal (com os valores passados das variáveis, no total de 12 meses) na análise, obtendo desta forma um total de 5390 variáveis.

Os 5 trabalhos a seguir utilizaram dados demográficos e histopatológicos: (SAFIYARI; JAVIDAN, 2018), (ALJAWAD et al., 2017), (KARHADE et al., 2018), (POURHOSSEINGHOLI; KHEIRIAN; ZALI, 2017), (SAILER et al., 2015). Destaque para (KARHADE et al., 2018), que implementou o algoritmo desenvolvido em uma página *web*.

Strunkin, Namee e Kelleher (2012), utilizou dados demográficos, clínicos, histopatológicos além de dados não-estruturados (texto livre escrito pela equipe médica sobre o estado de cada paciente).

Os 4 trabalhos a seguir utilizaram apenas dados genéticos dos pacientes: (GAN; ZHENG; WANG, 2014), (EYAL; LAST; RUBIN, 2015), (ZHANG; TANG; WANG, 2017) e (FU; CHENG; DING, 2018). Destaque para (GAN; ZHENG; WANG, 2014), que utilizou uma base com mais de 11 mil genes.

Os 4 trabalhos a seguir utilizaram apenas dados histopatológicos pacientes: (LISBOA et al., 2008), (JAYASURYA et al., 2010), (SETHI, 2018) e (WU et al., 2012).

## 3.2 CONSIDERAÇÕES FINAIS

Conforme revisão sistemática que realizamos (constante no apêndice A deste trabalho), não havia trabalhos publicados em língua inglesa utilizando a AGA em conjunto com AM para realizar predição de desfechos de tratamento oncológico. O que existe na literatura é a utilização de domínios específicos para se realizar predições.

No entanto, tivemos conhecimento da dissertação de mestrado de Sena Sena (2016), que posteriormente apresentou os resultados sob forma de publicação (SENA et al., 2019), e que de forma pioneira utilizou a AGA para prever tais desfechos. Algumas características do trabalho apresentado:

- 608 Pacientes na base de dados.
- Utilizou-se algoritmos sem otimização, em seus valores padrão.
- Não foi descrito o design da rede neural utilizada.
- Foi utilizada a métrica AUC ROC para decisão, tendo obtido 0,833 utilizando o algoritmo Naive Bayes.

Dante destas questões, e de uma base de dados 3x maior, entendemos ser relevante o reestudo do tema em questão por outro pesquisador. Outro fato importante que detalha-se, no nosso trabalho, todo o pré-processamento realizado, de forma que possa ser objeto de análise e critica de outros pesquisadores.

## 4 METODOLOGIA

### 4.0.1 IMPLEMENTAÇÃO UTILIZADA

Este capítulo apresenta as fases da pesquisa realizadas no contexto deste trabalho.

### 4.1 FASES DA PESQUISA

Para realização deste trabalho, realizamos duas fases distintas, que serão detalhadas em seguida, são elas:

- Revisão bibliográfica exploratória.
- Análise e tratamento dos dados.
- Ciclos de *Design Science Research* para elaboração da proposta e sua avaliação.

#### 4.1.1 REVISÃO BIBLIOGRÁFICA EXPLORATÓRIA

De acordo com Creswell (2014), a revisão bibliográfica é uma importante ferramenta de pesquisa, pois auxilia no entendimento sobre o tema e na avaliação de problemas de pesquisas existentes, ajudando o pesquisador a formular e definir um escopo para determinada área de interesse. Ela é muito útil para a descoberta de informações relevantes relacionadas ao problema de pesquisa e, a partir disso, planejar novas possibilidades de resolução do problema.

No escopo dessa pesquisa, a revisão teve um caráter exploratório, realizada nas principais bases eletrônicas disponíveis para a comunidade profissional e científica, tais como *IEEE*, *ACM*, *Science Direct* e *Google Scholar*. Essa primeira fase da pesquisa foi importante na investigação dos conceitos envolvendo o uso de Big Data e aprendizado de máquina no contexto da saúde, além de metodologias para lidar com problemas de aprendizagem de máquina voltados para classificação. Essa revisão também serviu de base para a construção dos capítulos que abordam a introdução, o referencial teórico e trabalhos relacionados.

### 4.2 ANÁLISE E TRATAMENTO DOS DADOS

De acordo com Pyle (1999) a etapa de tratamento dos dados (ou pré-processamento) é tida como uma das tarefas mais trabalhosas e demoradas no desenvolvimento de modelos de AM, tomando aproximadamente 80% do tempo despendido. No entanto, ela garante o

bom desempenho dos modelos de AM que serão desenvolvidos posteriormente. Dedicase o Capítulo 5 para este fim.

## 4.3 DESIGN SCIENCE RESEARCH (DSR)

Esta seção apresenta uma visão geral do *DSR* conforme protocolo descrito por Wieringa (2014), bem como a definição de ciclos, artefatos e métodos de avaliação que são utilizados pelo *DSR*, além de uma representação da aplicação do *DSR* no contexto desta pesquisa.

### 4.3.1 VISÃO GERAL DO DSR

Para Wieringa (2014), a DSR é um tipo de metodologia de pesquisa que enfatiza a conexão entre conhecimento e prática, mostrando que é possível produzir conhecimento científico a partir de projetos que tenham alguma utilidade prática e, portanto, visa resolver dois tipos de problemas a saber: problemas práticos e problemas de conhecimento, detalhados abaixo:

- **Problemas práticos:** que demandam uma mudança no mundo real que melhor coincide com alguns objetivos dos *stakeholders* (partes interessadas), ou seja, o mundo real sofre alguma mudança para se adaptar a objetivos da humanidade.
- **Problemas de Conhecimento:** que por outro lado não demandam uma mudança no mundo, mas uma mudança em nosso conhecimento sobre o mundo, ou seja, algum conhecimento sobre o mundo real é obtido sem obrigatoriamente alterá-lo.

Ao utilizar DSR, o fator principal para guiar a pesquisa é um problema prático, que baseado nele surgirão outros problemas práticos e questionamentos relacionados ao conhecimento (WIERINGA, 2014). Justificando ainda a escolha, de acordo com Engström, Storey e Runeson (2020), a metodologia do *Design Science* é um paradigma para a condução e comunicação de pesquisa aplicada, como é o caso da engenharia de Software.

### 4.3.2 CICLOS DO DSR

Segundo Wieringa (2014), um projeto que utiliza Design Science como metodologia interage sobre as atividades de projetar e investigar, através de um grande ciclo chamado de ciclo de engenharia, que é um processo racional de solução de problemas que pode ser representado pela Figura 9, e consiste das seguintes etapas:

1. **Investigação do problema:** Tem o objetivo de verificar quais fenômenos devem ser melhorados e o porquê. Trata-se de uma questão de conhecimento pois está

diretamente ligada ao entendimento do problema. Esse problema pode ser um novo ou um subproblema resultante de um ciclo anterior.

2. **Design da solução:** Projetar um ou mais artefatos que poderiam tratar o problema. Não necessariamente a solução será projetada totalmente nessa etapa, pois frequentemente parte da solução é construída nas tarefas de validação e implementação.
3. **Validação da solução:** Avalia se os artefatos projetados tratariam o problema definido no ciclo. Consiste em verificar se o projeto irá realmente atender aos objetivos dos stakeholders, caso seja implementado.
4. **Implementação da solução:** Tem o objetivo de implementar um dos artefatos projetados no contexto do problema no mundo real. O significado da palavra implementação depende do artefato que está sendo projetado. Ou seja, trata-se da execução do que foi planejado nas etapas anteriores.
5. **Avaliação da implementação:** Qual o sucesso do tratamento? Os resultados são avaliados podendo ser o início de uma nova iteração no ciclo de engenharia.

O ciclo de engenharia é composto pelo *ciclo de design*, que compreende as três primeiras etapas de *investigação do problema*, *design da solução* e *validação da solução*, e mais duas etapas chamadas de *implementação da solução* e *avaliação da implementação* (WIERINGA, 2014). O ciclo de design (Figura 10) representa um conjunto de etapas relacionadas ao projeto de uma pesquisa utilizando DSR, que podem ser repetidas várias vezes por pesquisadores, com o objetivo de chegar a um projeto apropriado da solução, no qual resulta em uma solução validada do problema (WIERINGA, 2014). Neste pesquisa, realizaremos apenas os ciclos de design pois não implementamos o algoritmo desenvolvido no ambiente real para avaliação, o que completaria o ciclo de engenharia. Definiu-se esta implementação como um trabalho futuro.

Conforme a Figura 9, para cada etapa no ciclo de engenharia existem pontos de interrogação e exclamação que devem ser verificados em cada uma delas. Os pontos de interrogação indicam perguntas relacionadas ao conhecimento, e os pontos de exclamação indicam problemas de projeto (WIERINGA, 2014). Nas etapas de investigação do problema e de avaliação da implementação são levantados os mesmos pontos de questionamentos, porém com objetivos diferentes. No primeiro caso o objetivo é de se preparar para o projeto de uma solução, aprendendo mais sobre o problema a ser tratado, e no segundo caso o objetivo é avaliar uma solução após ter sido aplicada no contexto real do problema (WIERINGA, 2014).

Figura 9 – Ciclo de Engenharia. Os questionamentos indicam questões de conhecimento. As exclamações, problemas de design.



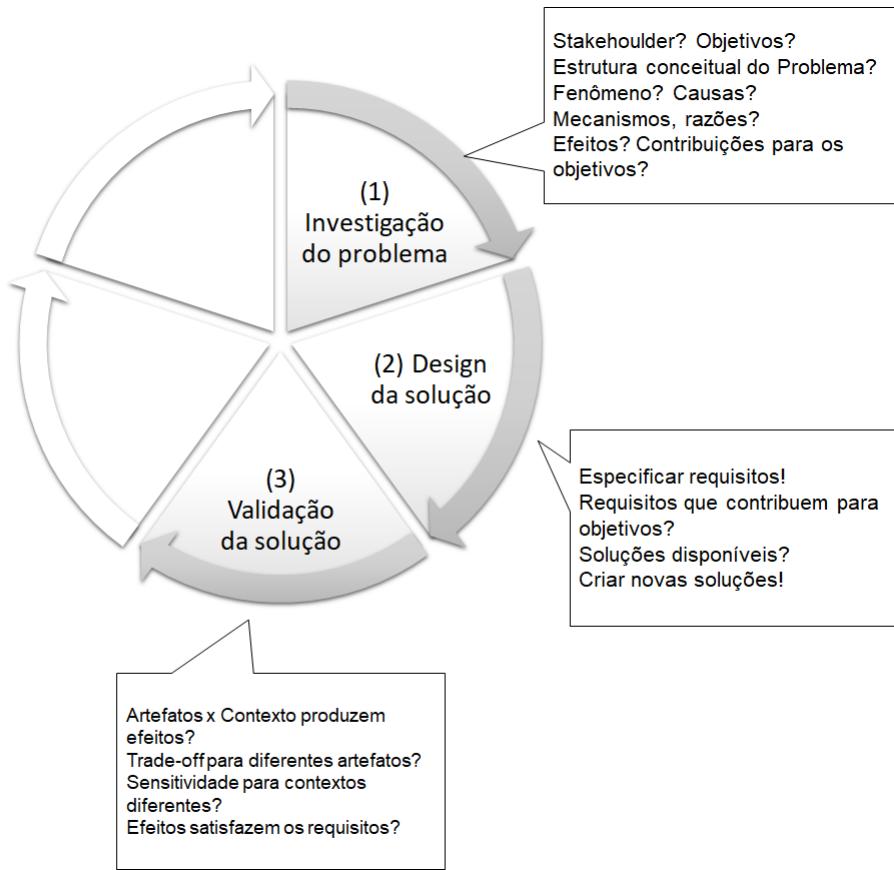
Fonte: Adaptado de Wieringa (2014)

#### 4.3.3 ARTEFATOS NO DSR

A metodologia de pesquisa DSR busca, a partir do entendimento de um problema, elaborar e avaliar artefatos que viabilizem a modificação de atividades ou situações para estados mais adequados ou pretendidos (DRESCH LACERDA, 2015). Para Hevner et al. (2004), um artefato é uma representação simbólica ou uma instanciação física. Já para Peffers et al. (2007), um artefato pode ser considerado qualquer coisa que foi projetada para atingir algum objetivo. Um artefato é algo criado por pessoas com algum propósito prático, e são utilizados ao projetar, desenvolver, implementar, manter e usar sistemas de informação e sistemas de software (WIERINGA, 2014). Alguns exemplos de artefatos projetados e estudados em sistemas de informação e pesquisa de engenharia de software citados por Wieringa (2014) são: algoritmos, métodos, notações, técnicas e até mesmo *frameworks* conceituais.

Para Hevner et al. (2004), artefatos de TI são amplamente definidos como construções (vocabulário e símbolos), modelos (abstrações e representações), métodos (algoritmos e práticas) e instanciações (sistemas implementados). No Quadro 6, estão relacionados os tipos de artefatos mais comuns elaborados por alguns pesquisadores como solução

Figura 10 – Ciclo de design. Os questionamentos indicam questões de conhecimento. As exclamações, problemas de design.



Fonte: Adaptado de Wieringa (2014)

tecnológica para algum tipo de problema (VAISHNAVI; KUECHLER, 2015).

Na seção 4.3.5 será informado qual tipo de artefato foi elaborado e avaliado nessa pesquisa.

Quadro 6 – Quadro com tipos de artefatos.

Tipo de Artefato	Descrição
Construções	Vocabulário conceitual de um domínio
Modelos	Conjunto de proposições ou declarações que expressam relacionamentos entre construções
FrameWorks	Guias reais ou conceituais para servir como suporte ou guia
Arquiteturas	Estruturas de alto nível de sistemas
Princípios de projeto	Princípios e conceitos chave para guiar o projeto
Métodos	Conjuntos de passos para executar tarefas
Instanciações	Implementações situada em certos ambientes que operacionalizam ou não construções, modelos, métodos e outros artefatos abstratos
Teorias de Projeto	Um conjunto prescritivo de instruções sobre como fazer algo para alcançar um determinado objetivo. Uma teoria geralmente inclui outros artefatos abstratos, como construções, modelos, frameworks, arquiteturas, princípios de projeto e métodos

Fonte: O autor. Adaptado de Vaishnavi e Kuechler (2015)

#### 4.3.4 MÉTODOS DE AVALIAÇÃO DE ARTEFATOS

Segundo Hevner et al. (2004), a avaliação é um componente crucial no processo de pesquisa, e a avaliação de um artefato de TI, que pode ser em termos de funcionalidade, integridade, consistência, acurácia, desempenho dentre outros atributos de qualidade, requer a definição de métricas apropriadas. No Quadro 7, são apresentados métodos e técnicas baseados em tipos de avaliações de artefatos propostos por Hevner et al. (2004). Na Seção 4.3.5 será informado qual método de avaliação foi utilizado nessa pesquisa.

Quadro 7 – Quadro com tipos, métodos e técnicas de avaliações de artefatos

Tipo de Avaliação	Métodos e Técnicas
<b>Observacional</b>	<b>Estudo de Caso:</b> Analisar o artefato de forma profunda no ambiente de negócios <b>Estudo de Campo:</b> Monitorar o uso do artefato em vários projetos
<b>Analítico</b>	<b>Análise Estática:</b> Analisar a estrutura do artefato para qualidades estáticas <b>Análise da Arquitetura:</b> Analisar a adaptabilidade do artefato na arquitetura técnica do sistema <b>Otimização:</b> Demonstrar as propriedades ótimas inerentes ao artefato ou fornecer os limites de otimização no comportamento do artefato <b>Análise Dinâmica:</b> Analisar o artefato durante o uso para avaliar suas qualidades dinâmicas
<b>Experimental</b>	<b>Experimento Controlado:</b> Analisar o artefato em ambiente controlado para verificar suas qualidades <b>Simulação:</b> Executar o artefato com dados artificiais
<b>Descriptivo</b>	<b>Argumento informado:</b> Utilizar informação de base de conhecimento (pesquisas relevantes, por exemplo) para construir um argumento convincente com relação a utilidade do artefato <b>Cenários:</b> Construir cenários detalhados em torno do artefato para demonstrar sua utilidade

Fonte: O autor

#### 4.3.5 APLICAÇÃO DO DSR NESSA PESQUISA

Na segunda fase dessa pesquisa, utilizou-se o DSR para a criação dos modelos de interesse da pesquisa. A fase foi importante para, a partir da área de pesquisa, investigar como poderia ser construído um modelo.

Segundo Wieringa (2014), os projetos de pesquisa que utilizam DSR não executam todo o ciclo de engenharia, restringindo-se apenas a execução do ciclo de design (figura 9) e a transferência de novas tecnologias para o mercado pode ser feita após o término do projeto de pesquisa, não fazendo parte do mesmo.

Como o objetivo geral dessa pesquisa é propor um modelo de classificação de pacientes oncológicos com relação ao risco de morte precoce, o tipo de artefato elaborado e avaliado

nos ciclos foi o **modelo**, conforme definido entre os tipos de artefatos descritos na seção 4.3.3.

- O primeiro ciclo, chamado de *seleção de features e do algoritmo de AM*, foi responsável por identificar as features relevantes e as features necessárias para o problema bem como avaliar o desempenho inicial dos 7 algoritmos de AM (Regressão Logística, Análise Discriminante Linear, K Vizinhos Próximos, Naive Bayes, Máquina de Vetores de Suporte, Perceptron Multicamadas e XGboost) utilizando as variáveis identificadas como necessárias como entrada. Ao fim do primeiro ciclo, foi possível obter a performance inicial de cada um dos algoritmos utilizados.
- No segundo e último ciclo, chamado de *otimização*, realizamos ajustes nos algoritmos utilizados no ciclo anterior a fim de melhorar o desempenho. Em seguida, o modelo será finalizado ao ponto de ser utilizado em uma futura aplicação.

Para avaliação, considerando as informações propostas por Hevner et al. (2004), apresentadas na Tabela 7, para o contexto desse trabalho, enquadram-se os métodos de avaliação **Experimental** e **Descriptivo** com as seguintes justificativas:

- **Experimental**

1. Dados históricos foram analisados e tratados para posterior uso em modelos de AM.
2. As variáveis mais significativas foram selecionadas para utilização no modelo.
3. Avaliou-se a precisão dos modelos testados utilizando-se a área sobre a curva ROC, curva precision-Recall e recall.

- **Descriptivo**

1. Nossos modelos foram comparados a modelos similares, de outros pesquisadores.

Segundo Chiavegatto et al. (2019), há diversas métricas para a avaliação do desempenho de modelos preditivos com base no paradigma de aprendizado supervisionado. Tais métricas buscam mensurar o desempenho das previsões decorrente do modelo ajustado, avaliando o quanto ele reproduz o valor observado para a resposta de interesse. Para modelos de predição por classificação, como o nosso caso, normalmente se realiza uma tabulação cruzada das classes observadas e preditas em uma matriz de confusão, apresentada na Figura 5.

A partir dessa matriz, diversas métricas podem ser calculadas, conforme foi detalhado na seção 2.2.8 do capítulo 2, onde também foi descrita a fundamentação teórica relacionada ao contexto da pesquisa.

# 5 ANÁLISE E TRATAMENTO DOS DADOS

Este capítulo inicia explicitando na Seção 5.1 algumas decisões importantes realizadas durante a pesquisa. Em seguida, na Seção 5.2, realiza-se a análise e pré-processamento dos dados do centro onde este estudo foi realizado (IMIP).

## 5.1 DECISÕES DA PESQUISA

Esta Seção descreve e justifica importantes decisões tomadas no contexto da pesquisa.

### 5.1.1 IMPLEMENTAÇÃO UTILIZADA

Toda a análise dos dados e modelos de AM desta dissertação foram realizadas utilizando a linguagem Python (2020) e além do Scipy (2020), que é um ecossistema de software de código aberto baseado em *Python* para matemática, ciência e engenharia e que conta com pacotes como *NumPy*, *pandas*, *Matplotlib*. Além deles, utiliza-se o Scikit-Cearn (2020), que é uma biblioteca de aprendizado de máquina de código aberto com diversos algoritmos prontos para uso.

*Python* é uma linguagem de programação de alto nível, interpretada, de script, imperativa, orientada a objetos, funcional, de tipagem dinâmica e forte. Foi criada por Guido van Rossum em 1991 e atualmente possui um modelo de desenvolvimento comunitário, aberto e gerenciado pela organização sem fins lucrativos Python Software Foundation (PSF).

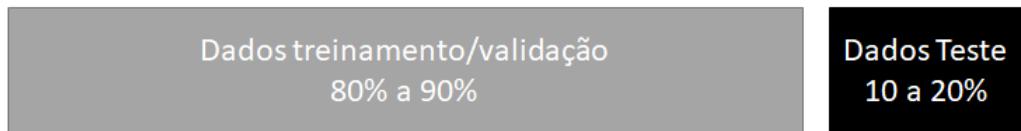
### 5.1.2 DIVISÃO DO DATASET

Neste trabalho, adotou-se a nomenclatura para divisão da base de dados conforme descrita por Bishop (1995) e Ripley (1996) e resumida no F.A.Q. SAS (2017), que divide a base de dados em treinamento/validação e teste, conforme descrito abaixo e na Figura 11:

- **Base de treinamento:** Base de dados utilizada para aprendizado dos modelos de AM, para ajuste dos parâmetros. A base de treinamento normalmente tem entre 80% e 90% dos dados disponíveis. Durante o treinamento dos modelos de AM, ela é dividida novamente, separando uma parte dos dados para a base de validação.
- **Base de validação:** Utilizada para avaliar a performance dos modelos durante o treinamento. É uma subdivisão da base de treinamento.
- **Base de teste:** Base de dados utilizada apenas para avaliação final do modelo. São dados não vistos pelo modelo durante o treinamento. O ideal é que sejam dados

novos, se não for possível, pode-se separar uma parte dos dados para este fim. Valores típicos para base de teste são entre 10% e 20% dos dados disponíveis.

Figura 11 – Exemplo de divisão dos dados durante o tratamento.



Fonte: O autor

A divisão acima é importante para evitar o problema do vazamento de dados (*data leakage*), conforme descrito na Seção 2.2.3. Adotaremos, neste trabalho, a divisão em 80% dos dados para treinamento e validação e os 20% finais para teste. A divisão será feita de forma aleatória.

#### 5.1.2.1 VARIÁVEL DE SAÍDA: DEFINIÇÃO DA CLASSE POSITIVA

Como o objetivo é prever a ocorrência de óbito em 6 meses a contar da data da realização da AGA, **seguiremos o padrão abaixo para as classes da variável de saída, motivo da saída.**

- **Classe positiva (1):** Óbito em menos de 6 meses.
- **Classe negativa (0):** Não ocorrência de óbito em 6 meses. Vida.

Os modelos de AM avaliados seguirão este padrão na realização de predições.

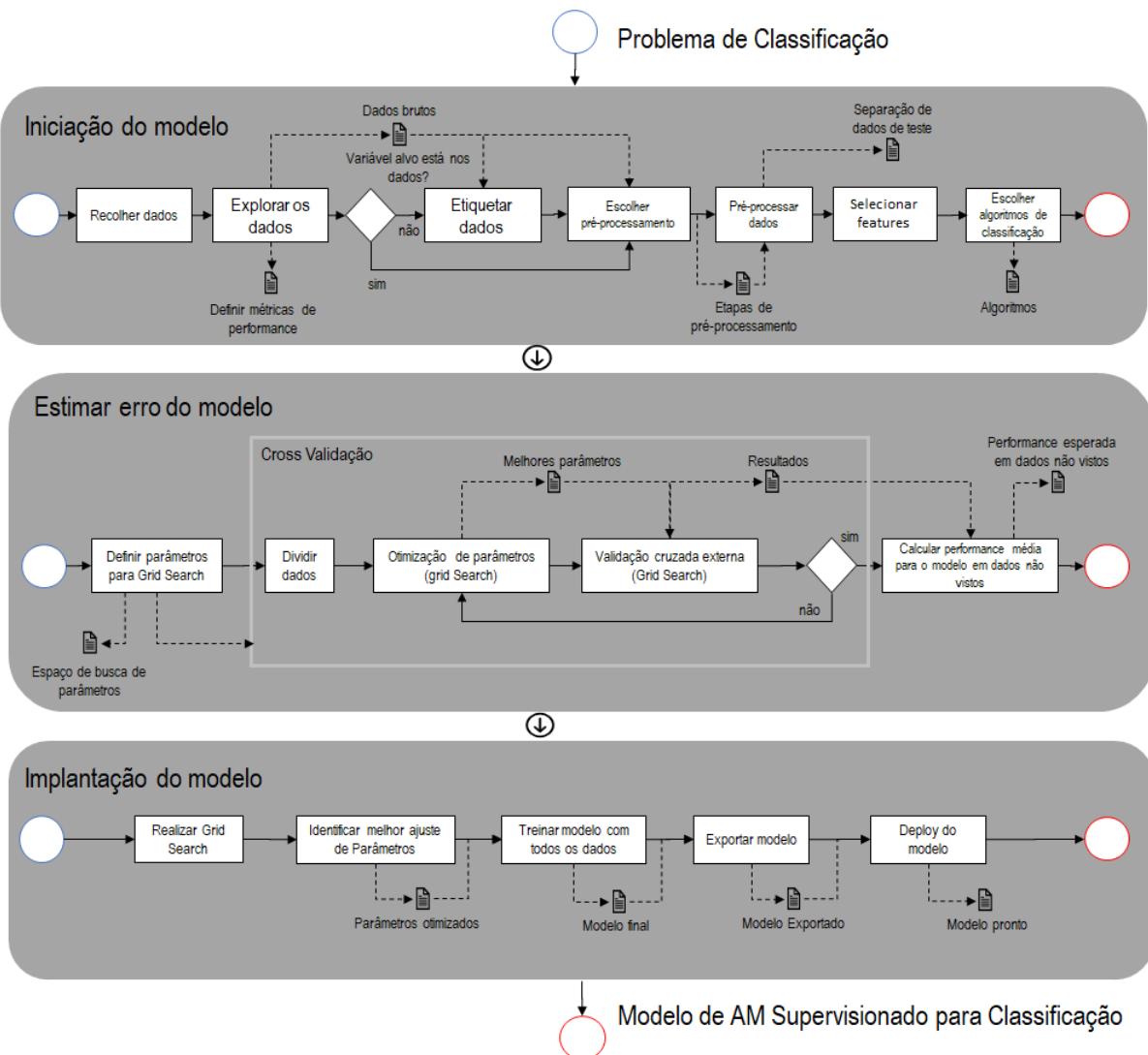
#### 5.1.3 MÉTRICAS UTILIZADAS

O nosso objetivo é reduzir o numero de falsos negativos, idealmente a zero. Um falso negativo, neste contexto, significa um paciente que morreria em menos de 6 meses, mas foi classificado incorretamente. Portanto, as métricas que iremos utilizar são *AUC-Precision-Recall*, *recall*, *f1-score* e *AUC-ROC*. As três primeiras, conforme descrito no capítulo 2, são as métricas mais relevantes quando avaliamos a classe positiva (minoritária) em uma base de dados desbalanceada. A *AUC-Precision-Recall* será utilizada para classificar a performance dos algoritmos e as demais são informativas, com intuito de comparar com outros trabalhos que utilizaram estas métricas.

### 5.1.4 PROCESSO DE DESENVOLVIMENTO DE MODELO DE AM PARA CLASSIFICAÇÃO

Embora existam diversos modelos diferentes de processos de mineração de dados em geral, conforme pesquisa realizada por Kurgan e Musilek (2006), as abordagens existentes não foram desenvolvidas especificamente para problemas de classificação como o modelo proposto por Hirt, Niklas e Satzger (2017) e detalhado na Figura 12. Utilizou-se este modelo na pesquisa.

Figura 12 – Processo de desenvolvimento de modelos de AM para problemas de classificação.



Fonte: Adaptado pelo autor. Original de Hirt, Niklas e Satzger (2017)

O modelo proposto descreve três macro etapas de desenvolvimento: iniciação, estimativa de erros e implantação. Ao final, terminamos com um modelo de classificação AM pronto para ser implantado.

Porém o modelo proposto por Hirt, Niklas e Satzger (2017) não define como deve ser a

escolha do algoritmo de AM utilizado (etapa *Escolher algoritmo classificação* da fase de iniciação).

O processo inicia-se com a definição do problema em estudo, identifica-se qual a variável objetivo (variável de saída) a ser prevista. Na primeira fase, chamada *iniciação do modelo*, o espaço do problema é explorado (*explorar dados*) para se obter *insights* sobre os dados. Métodos, como uma análise exploratória de dados, são úteis para encontrar padrões significativos e identificar os dados relevantes que podem ser processados posteriormente. Além disso, selecionam-se as métricas de desempenho que serão posteriormente utilizadas para validar e testar o modelo. As métricas são selecionadas em função da configuração do problema específico, por exemplo, refletindo os impactos dos erros de classificação.

Em seguida, coletam-se os dados brutos (etapa *recolher dados*). Caso o atributo alvo ainda não tenha sido definido, é necessário um processo de rotulagem manual (*rotular dados*). Nesse caso, para garantir a exatidão dos mesmos, é aconselhável categorizá-los por mais de um avaliador humano para minimizar o viés de classificação manual. Em seguida, estes dados são limpos, estruturados e posteriormente pré-processados (etapa *pré-processar dados*). Uma parcela deles é removida para posterior avaliação de desempenho (*separação de dados de teste*). A única etapa restante é a escolha de um algoritmo de classificação (*escolher algoritmo de classificação*).

A segunda fase chama-se *estimar erro do modelo*, que tem por objetivo estimar o desempenho esperado do modelo em dados não vistos. Com o objetivo é construir um modelo sólido, precisa-se considerar diferentes parâmetros dos algoritmos, que caracterizarão o modelo de AM. Diversos parâmetros internos dos algoritmos podem ser ajustados para uma melhor performance. Para cada parâmetro escolhe-se um intervalo de valores (espaço de busca) para que, em seguida, a performance do algoritmo seja avaliada para todas as combinações de valores possíveis (etapa *definir parâmetros para Grid Search*).

Em seguida, dividem-se os dados (*dividir dados*) para serem tratados em validação cruzada. O modelo é treinado e validado para diferentes conjuntos de parâmetros para sua adequação ao problema em questão. Várias técnicas para ajuste de parâmetros como busca em grade (*grid search*), otimização bayesiana, otimização baseada em gradiente ou seleção aleatória podem ser usadas, conforme Hirt, Niklas e Satzger (2017). Em última análise, as métricas previamente definidas determinam os parâmetros de melhor desempenho. Após a iteração através de todas as possibilidades (no caso de uma busca em grade), utilizam-se os parâmetros de melhor desempenho (em média) para treinar todo o conjunto e validá-lo em dados não vistos (etapa *Calcular performance média para o modelo em dados não vistos*). Após esta etapa, estima-se a performance do modelo em dados novos.

A fase final (*implantação do modelo*) é responsável por preparar o modelo para implantação em um sistema de informação. Como os dados são sempre valiosos e na maioria dos casos escassos, utilizam-se o conjunto completo de dados para treinar o modelo final de

classificação de AM utilizando os parâmetros previamente selecionados. Primeiramente, realiza-se outra busca em grade (*grid search*) com o mesmo espaço de busca de parâmetros a partir da estimativa de erros, porém com todos os dados disponíveis. Identifica-se a combinação de parâmetros que alcançam os melhores resultados em relação à nossa métrica de desempenho pré-definida. Em seguida o modelo é treinado novamente, utilizando todos os dados. Em seguida, uma exportação do modelo final é necessária para salvar o estado do modelo. Agora, o objeto serializado pode ser incluído em um *workflow*, como um serviço web conectado, para prever o valor alvo dos dados novos e recebidos.

## 5.2 ANÁLISE DOS DADOS

Para realizar este estudo, contamos com os dados provenientes do IMIP que, realiza mais de 3 mil atendimentos oncológicos por mês sendo umas das poucas instituições, entre públicas e privadas, que realiza os exames que compõem a AGA e os digitaliza, mesmo que de forma manual, para posterior acompanhamento e utilização em pesquisa.

Todo novo paciente oncológico idoso (acima de 60 anos) admitido no IMIP para tratamento, realiza as avaliações que compõem a AGA que se traduz em um conjunto de 107 variáveis ou *features*, resumidas no quadro 8. Este estudo é retrospectivo, que é aquele em que todos os dados já foram coletados, para realizar o prognóstico com relação ao desfecho (óbito ou não) do paciente após 6 meses de acompanhamento.

Esta fase tem como objetivo realizar a análise e o tratamento dos dados obtidos no IMIP para posterior utilização nos modelos de AM. Os modelos e técnicas utilizados estão descritos no capítulo de revisão bibliográfica.

Os dados, por conterem informações pessoais, são sigilosos e só foram fornecidos mediante assinatura de termo de responsabilidade nosso (autor e orientadora). Eles são utilizados em diversas pesquisas no IMIP, incluindo a nossa. Fazem parte de um projeto de pesquisa maior, aprovado pelo Comitê de Ética em Pesquisa em Seres Humanos (CEP) do IMIP sob o número **58298316.5.0000.5201**. Importante ressaltar que, dado o caráter universal da AGA e das variáveis utilizadas, esta análise e modelo de AM desenvolvido se aplicam a qualquer centro, com pequenas alterações.

O objetivo nesta fase é:

- Como preparar os dados fornecidos pelo IMIP para uso nos modelos de AM, de forma a maximizar a performance dos mesmos?

Para responder a esta pergunta iniciamos com a análise exploratória dos dados (Seção 5.2.1) e, em seguida, pelo tratamento dos dados de forma a maximizar a performance dos modelos de AM.

Quadro 8 – Quadro resumo com a quantidade de variáveis da AGA por cada categoria.

Avaliação	Descrição	Total de variáveis
QLQ-C30	Questionário para avaliar a qualidade de vida do paciente com câncer. É composto de 30 questões.	30
Mini Avaliação Nutricional(MAN)	Avaliação do estado nutricional do paciente	12
Sócio Demográfica	Dados gerais do Paciente	18
Índice de Charlson	Avalia a presença de comorbidades no paciente	16
Hábitos pregressos	Relacionados ao histórico de consumo de Álcool e Cigarro	10
Exames laboratoriais	Hemoglobina, contagem de leucócitos, granulócitos, plaquetas e creatinina	5
Antecedente Vacinal	Histórico vacinal para Tétano, Pneumonia, Influenza e Hepatite-B	4
Timed up and go	Avaliação de mobilidade	3
Diagnóstico	Código CID C 10 do câncer diagnosticado além da informação da presença ou não de metástase. Além da localização da metástase	2
Polifarmácia	Avalia a quantidade de medicamentos de uso rotineiro utilizados pelo paciente	1
Escala de Performance de Karnofsky	É uma medida relacionada à tentativa de quantificar o bem-estar geral dos pacientes	1
Mini Mental	Avaliação do estado cognitivo do paciente	1
GDS escala de depressão	Escala de depressão geriátrica	1
PPS	Escala de performance Paliativa	1
Índice de KATZ-AVD	Avaliação de funcionalidade em atividades da vida diária	1
Avaliação Física	Classificação de atividade Física	1
Total		107

Fonte: O autor

Utilizaremos o Workflow (Figura 12) adaptado de Hirt, Niklas e Satzger (2017) para nos guiar durante a elaboração do nosso modelo de AM para o nosso problema de classificação. Durante esta etapa, nos ateremos a primeira fase, *inicialização do modelo*, até a etapa de *pré-processamento dos dados*.

As etapas de pré-processamento estão definidas na Seção 2.2.1.

### 5.2.1 ANÁLISE EXPLORATÓRIA DOS DADOS

O IMIP nos forneceu um conjunto de 10 arquivos no formato .xlsx com dados referentes a 1653 pacientes e 136 variáveis. A lista completa das variáveis está no apêndice B.

Segue uma breve análise dos pacientes que constam na base de dados do IMIP:

- **Estado:** 99,1% deles são de Pernambuco.
- **Sexo:** 54,9% de homens e 45,1% de mulheres.
- **Diagnóstico principal:** 32,2% são de câncer de próstata, seguidos de mama (14,9%), estômago (6,5%), Cólono (4,6%), Colo do útero (4,2%), endométrio (4,1%), bexiga (4,0%), esôfago (3%), reto (3%) e outros (18,4%).
- **Idade:** Média (70,3), mediana (70), Q1 (65), Q3 (76), Desvio-padrão (8,97).
- **Motivo da saída:** 84,6% de pacientes vivos e 15,4% de óbitos. Esta é a variável de saída. Verifica-se que está desbalanceada.

São 136 variáveis listadas, onde 103 são categóricas, 33 são numéricas, além do registro do paciente e da variável de saída (Motivo da Saída).

Os dados obtidos são de pacientes admitidos entre 2015 e 2019. A base de dados já está rotulada, quer dizer, todos os pacientes já estão com o desfecho definido, se vivo ou não ao final de 6 meses de acompanhamento.

Cinquenta e três registros foram removidos pois eram duplicados ou de pacientes sem número de registro, restando neste ponto, 1600 registros.

### 5.2.2 PRÉ-PROCESSAMENTO DOS DADOS

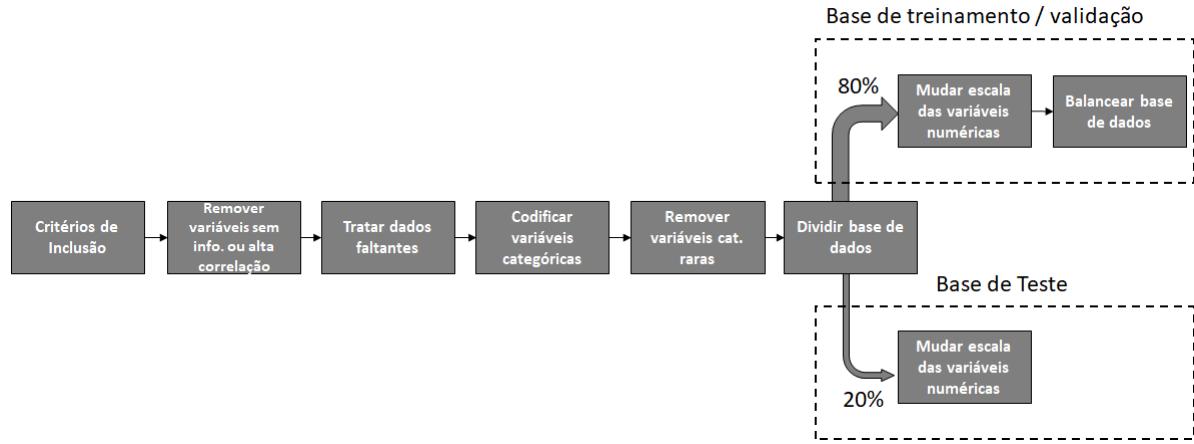
Nesta Seção, executam-se as etapas de pré-processamento dos dados, definidas na Seção 2.2.1 e mapeadas na Figura 13.

#### 5.2.2.1 CRITÉRIOS DE INCLUSÃO

Os critérios de inclusão na pesquisa clínica estão descritos no ANEXO I, onde consta a ficha completa com formulário de admissão incluso, onde estão descritos todos os critérios médicos para admissão. Os critérios para inclusão na base de dados da pesquisa são:

1. Estar na base de dados do IMIP (ie. ter atendido a todos os critérios médicos estabelecidos)
2. Ter completado 6 meses de acompanhamento, ou falecido antes.

Figura 13 – Etapas de pré-processamento definidas e executadas.



Fonte: O autor.

Como nosso objetivo é classificar os pacientes entre baixo e alto risco de morte em 6 meses a contar da admissão (preenchimento do formulário e execução dos testes da AGA), manteve-se na base quem efetivamente cumpre o período ou faleceu antes.

### 5.2.3 REMOÇÃO DE VARIÁVEIS SEM INFORMAÇÃO OU COM ALTA CORRELAÇÃO

As variáveis do Quadro 9 abaixo serão removidas da base de dados, conforme discussão abaixo:

Quadro 9 – Variáveis removidas.

Atributo	Tipo	Range
Estado	Categórica	-
Município	Categórica	-
Bairro	Categórica	-
Diagnóstico Principal	Categórica	-
Karnofsky	Categórica	0 .. 100
Metastase-Sitios-Dsc	Categórica	-
Anteced-Hist-Quedas-Qtas	Categórica	maior ou igual a 0
Etil-Frequencia	Categórica	Texto com frequência de ingestão de álcool
Etil-Quanto-Tempo	Categórica	Texto com histórico temporal
Etil-Tipo-Bebida	Categórica	Texto com o tipo da bebida frequentemente ingerida
Etil-Ja-Bebeu-Em-Que-Qtde	Categórica	Texto com alegada frequência de ingestão de álcool
Etil-Ja-Bebeu-Tipo-Bebida	Categórica	Texto com o tipo da bebida frequentemente ingerida

Fonte: O autor

- **Estado, Município, e Bairro:** Removidas por, claramente, não terem relação com a variável de saída.
- **Diagnóstico principal:** Removida pois a variável CID-10C tem a mesma informação codificada no CID-10, Classificação Internacional de Doenças, publicada pela OMS.
- **Karnofsky:** O índice Karnofsky é uma medida relacionada à tentativa de quantificar o bem-estar geral dos pacientes. Pode ser utilizada para determinação da possibilidade de receber quimioterapia, da necessidade de ajuste de doses destas medicações, entre outras finalidades. O índice PPS (*Paliative Performance Index*) é derivado do Karnofsky, porém com a finalidade de ser aplicado em pacientes em cuidados paliativos. O fator de correlação Spearman entre as duas apresentou-se superior a 0.90, indicando alta correlação, segundo Mukaka (2012). Diante disto, manteremos o índice PPS Kenis e Wildiers (2011).
- **Metastase-Sitios-Dsc:** A variável **Metastase-Sitios** tem a mesma informação de forma categórica (1- Óssea, 2- Pulmonar, 3 - SNC, 4 - Hepática, 5 - Outras , 9 - Sem ocorrência de metástase).
- **Anteced-Hist-Quedas-Qtas:** Apresenta 81% de dados faltantes e, por este motivo, será removida. Existe a variável **Anteced-Hist-Quedas** que apresenta a informação se há ou não histórico de quedas recentes, porém sem a contagem.
- **Etil-Frequencia e Etil-Ja-Bebeu-Em-Que-Qtde:** Removida por apresentar 91% de dados faltantes e os remanescentes possuem informações pouco precisas.

- **Etil-Quanto-Tempo:** Removida por apresentar 91% de dados faltantes e os remanescentes possuem informações pouco precisas.
- **Etil-Tipo-Bebida e Etil-Ja-Bebeu-Tipo-Bebida:** Removidas por apresentarem 91% de dados faltantes e os remanescentes possuem informações pouco precisas.

#### 5.2.3.1 TRATAMENTO DE DADOS FALTANTES

A estratégia utilizada para tratamento de dados faltantes para variáveis categóricas foi utilizar o valor da classe mais frequente, para cada variável. Para as variáveis numéricas, adotamos o uso da mediana, ou seja, aos valores faltantes será atribuído o valor da mediana dos valores da variável em questão. Nas Tabelas 1 e 2 encontra-se o quantitativo de dados faltantes para as variáveis categóricas e o quantitativo de dados faltantes para as variáveis numéricas, respectivamente.

Tabela 1 – Quantitativo de dados faltantes em variáveis categóricas.

Variável	Número de Dados Faltantes	Valor percentual
CID10	26	1,63%
Time-Up-Go-Classi-Mobil	6	0,38%
Time-Up-Go-Classi-Ris-Qued	6	0,38%
ATV-Fisica-Classif	4	0,25%
MAN-AvG-H	3	0,19%
MAN-AvG-I	3	0,19%
MAN-AvG-L	3	0,19%
Q.02	2	0,13%
Q.03	2	0,13%
Q.04	2	0,13%
1-Infarto Mioc	2	0,13%
1-Insuf Card Cong	2	0,13%
1-Doenca Vasc Perif	2	0,13%
1-Doenca Cereb Vasc	2	0,13%
1-Doenca Pulm Cron	2	0,13%
1-Doenca Tecido Conj	2	0,13%
1-Diabete Leve S Complic	2	0,13%
1-Ulcera Peptica	2	0,13%
2-Hemiplegia	2	0,13%
2-Diabete C Complic	2	0,13%
2-Doenca Renal Sev Mod	2	0,13%
3-Doenca Fig Sev Mod	2	0,13%

**Tabela 1** continuação da página anterior

Variável	Número de Dados Faltantes	Valor percentual
6-Tumor Malig Metas	2	0,13%
6-SIDA	2	0,13%
0-HAS	2	0,13%
MAN-TR-A	2	0,13%
MAN-TR-B	2	0,13%
MAN-TR-C	2	0,13%
MAN-TR-D	2	0,13%
MAN-TR-E	2	0,13%
MAN-AvG-K-4-Res	2	0,13%
MAN-Aval-Est-Nut-Res	2	0,13%
MAN-AvG-G	2	0,13%
MAN-AvG-J	2	0,13%
MAN-AvG-K-1	2	0,13%
MAN-AvG-K-2	2	0,13%
MAN-AvG-K-3	2	0,13%
MAN-AvG-M	2	0,13%
MAN-AvG-N	2	0,13%
MAN-AvG-O	2	0,13%
MAN-AvG-P	2	0,13%
Metastase-Sitios	1	0,06%
Tabag-Q.02	1	0,06%

Fonte: O autor

**Tabela 2 –** Variáveis numéricas faltantes.

Variável	Número de Dados Faltantes	Valor percentual
5-Polifarmacia-Medicamento	246	15,38%
Idade(anos)	7	0,44%
Peso	6	0,38%
Altura	5	0,31%
Q.05	2	0,13%
MAN-Escore-TR	2	0,13%
MAN-Aval-Global	2	0,13%
MAN-Escore-Tot	2	0,13%

**Tabela 2** continuação da página anterior

Variável	Número de Dados Faltantes	Valor percentual
MAN-Aval-Est-Nutric	2	0,13%
MAN-TR-F	2	0,13%
MAN-AvG-Q	2	0,13%
MAN-AvG-R	2	0,13%

Fonte: O autor

### 5.2.3.2 CODIFICAÇÃO DAS VARIÁVEIS CATEGÓRICAS

Aplicou-se o método de codificação de variáveis categóricas conhecido como *one-hot encoding*, conforme descrito na Seção 2.2.4, resultando em 408 variáveis.

### 5.2.3.3 REMOÇÃO DE VARIÁVEIS CATEGÓRICAS RARAS

Para evitar o *overfitting*, removeremos do dataset as variáveis categóricas raras, ou seja, com poucas ocorrências. Conforme Luo et al. (2016), uma abordagem conservadora é remover do dataset as variáveis com menos de 10 ocorrências. Sendo assim, removeu-se do dataset 57 variáveis categóricas, presentes na tabela 3, que apresentaram esta característica.

Tabela 3 – Variáveis categóricas codificadas removidas do dataset por apresentarem menos de 10 ocorrências.

Variável Codificada	Número de ocorrências
CID10_6.0	1
CID10_7.0	1
CID10_26.0	1
CID10_27.0	1
CID10_30.0	1
CID10_37.0	1
CID10_38.0	1
CID10_45.0	1
CID10_46.0	1
CID10_60.0	1
CID10_66.0	1
CID10_76.0	1
CID10_78.0	1
CID10_82.0	1

**Tabela 3** continuação da página anterior

Variável Codificada	Número de ocorrências
CID10_86.0	1
CID10_97.0	1
CID10_98.0	1
QLQ30-29_0	1
CID10_0.0	2
CID10_4.0	2
CID10_12.0	2
CID10_35.0	2
CID10_73.0	2
CID10_74.0	2
Q.02_88.0	2
QLQ30-30_0	2
CID10_1.0	3
CID10_28.0	3
CID10_68.0	3
06-Cor Pele_5	3
06-Cor Pele_77	3
Q.04_4.0	3
CID10_31.0	4
CID10_52.0	4
Q.01_SUPERIOR INCOMPLETO	4
CID10_2.0	5
CID10_11.0	5
CID10_48.0	5
CID10_72.0	5
06-Cor Pele_88	5
Tabag-Q.01_2	5
CID10_21.0	6
CID10_24.0	6
CID10_71.0	6
CID10_83.0	6
CID10_85.0	6
CID10_10.0	7
CID10_19.0	7
CID10_44.0	7

**Tabela 3** continuação da página anterior

Variável Codificada	Número de ocorrências
CID10_51.0	7
6-SIDA_1.0	7
Tabag-Q.01_0	7
Tabag-Q.02_0.0	7
Etil-Consume-Beb-Alcool_0	7
Etil-Ja-Bebeu_0	7
CID10_80.0	8
1-Doenca Tecido Conj_1.0	9

Fonte: O autor.

#### 5.2.3.4 DIVISÃO DOS DADOS ENTRE TREINAMENTO/VALIDAÇÃO E TESTE

Conforme decisão da documentada na Seção 5.1.2, adotou-se a divisão em 80% dos dados para treinamento e validação e os 20% finais para teste, que foi realizada de forma aleatória. Ao todo são 1600 registros sendo 1280 (80%) utilizados na base de treinamento e validação, enquanto os 320 restantes foram utilizados na base de teste. Ambas as bases contam com 349 variáveis além da variável de saída, *motivo da saída*. Além disto, removeu-se o *Registro* do paciente com duplo objetivo: garantir o sigilo e prevenir possível vazamento de dados.

#### 5.2.3.5 MUDANÇA DE ESCALA NAS VARIÁVEIS NUMÉRICAS

Após a separação entre as bases de treinamento/validação e teste, realizou-se a transformação Z-score para as variáveis numéricas de ambas as bases, separadamente.

#### 5.2.3.6 BALANCEAMENTO DA BASE DE DADOS

Dos 1600 registros da nossa base de dados completa, 1340 pacientes tiveram o desfecho *vida* após 6 meses, enquanto 240 vieram a *óbito*. Uma proporção de 15% de óbitos, que foi mantida após a divisão da base em treinamento/validação e teste, devido a natureza aleatória da separação. Devido ao desbalanceamento, realizou-se, na base de treinamento/validação, a criação de registros adicionais por meio do algoritmo SMOTE-NC, apresentado na Seção 2.2.6. Após o isto o *dataset* restou com igual proporção de desfechos, ou seja, com 50% de amostras com pacientes que sobreviveram e 50% com pacientes que vieram à óbito em um total de 2194 registros (ou 1097 em cada classe). A base de dados de teste contou 320 Registros e, como não foi utilizada para treinamento, não houve necessidade de ser balanceada.

O objetivo desta Seção foi apresentar o pré-processamento de dados, seguindo o processo de desenvolvimento descrito por Hirt, Niklas e Satzger (2017). A próxima etapa seria selecionar as variáveis para, em seguida, escolher o algoritmo de AM a ser utilizado. Porém, como definiremos o algoritmo de AM a ser utilizado por meio de ciclos de DSR, considerou-se importante criar um modelo para ser uma referência comparativa com os próximos a serem desenvolvidos.

### 5.3 MODELO DE REFERÊNCIA

Para validar o tratamento de dados realizados, criamos um classificador utilizando o algoritmo da Regressão Logística, a primeira escolha para problemas de classificação conforme Brownlee (2019). Este classificador será utilizado como referência de performance para os que serão implementados em seguida.

O modelo de Regressão Logística foi treinado sem nenhuma otimização. Informamos apenas que a métrica de otimização seria '*f1*'. Além disso, o modelo foi treinado na base de treinamento/validação utilizando a validação cruzada (com  $K = 10$ ).

Em seguida, o desempenho foi avaliado na classificação da base de teste. O resultado está na Tabela 4 e nos gráficos da Figura 14.

Verifica-se que o classificador apresenta um desempenho ruim (0,68 para AUC ROC e 0,37 para *AUC Precision-Recall*), conforme classificação proposta por Kumari e Kishore (2018).

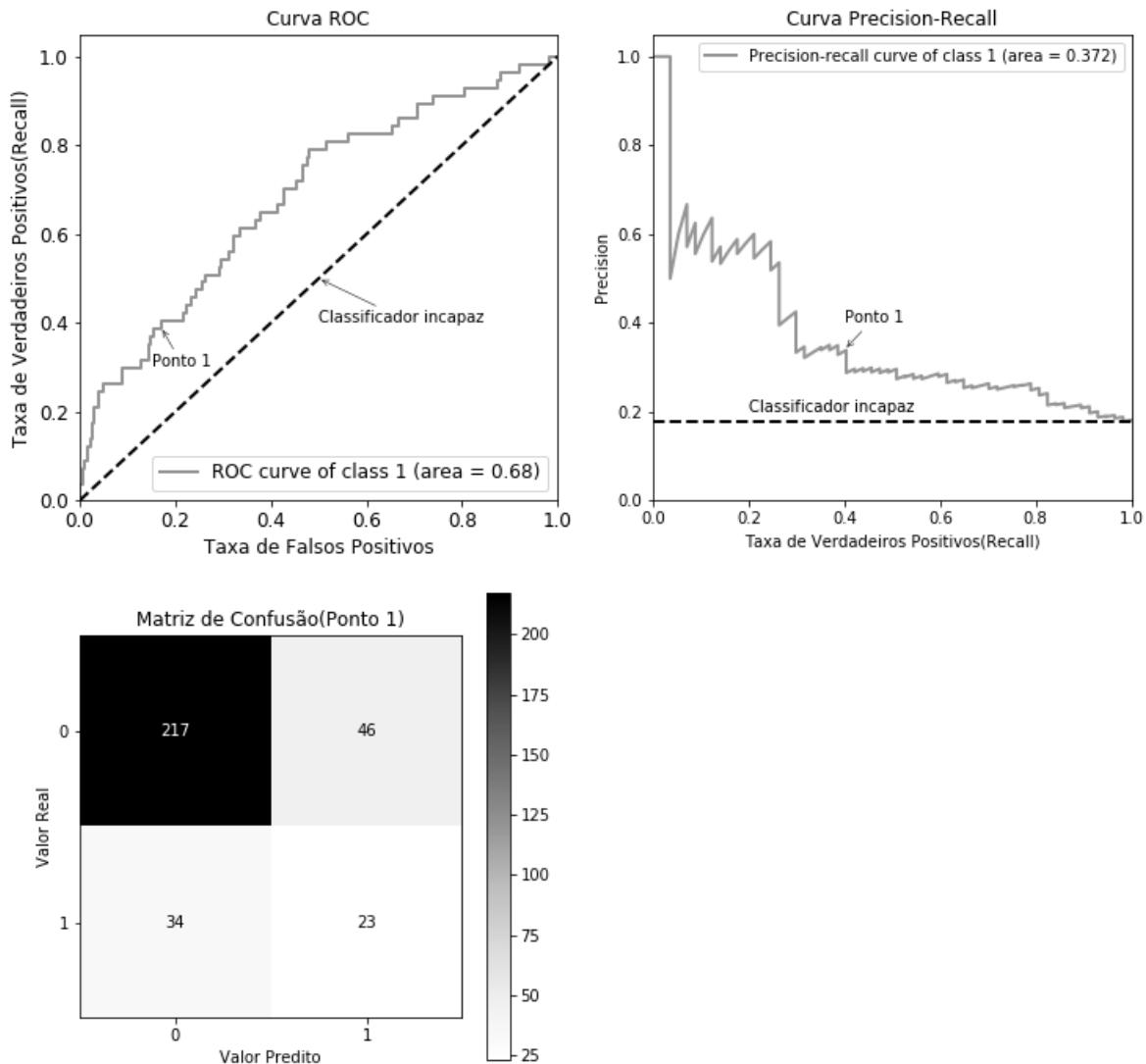
Da matriz de confusão (Figura 14), verifica-se que 23 pacientes de alto risco, que vieram a óbito em menos de 6 meses, foram classificados corretamente como VP, porém outros 34 pacientes não foram detectados no classificador. O objetivo, no nosso contexto, é detectar o máximo possível desses pacientes. Os demais 46 pacientes que não são do grupo de risco mas que foram erroneamente classificados como tal devem ser minimizados, mas não são o foco principal.

Tabela 4 – Métricas obtidas no classificador de referência.

Algoritmo	Precisão	Recall	Taxa Falsos Positivos	AUC ROC	AUC Precision-Recall	F1 score
Regressão Logística (Referência)	0,333	0,404	0,175	0,683	0,372	0,365

Fonte: O autor

Figura 14 – Classificador de referência.



Fonte: O autor

## 5.4 CONCLUSÃO

Neste capítulo, realizamos a análise exploratória e pré-processamento dos dados brutos obtidos no IMIP. Em seguida, treinamos um classificador para ser utilizado como referência para comparar com outros modelos que serão desenvolvidos em seguida por meio de DSR. O capítulo seguinte retoma o desenvolvimento do modelo de AM conforme descrito por Hirt, Niklas e Satzger (2017), realizando a seleção de variáveis.

# 6 CICLO 1 - SELEÇÃO DE FEATURES E DO ALGORITMO DE AM

Este capítulo tem como objetivo apresentar o primeiro ciclo de design, conforme identificado no capítulo 4. Neste ciclo, realizaremos a identificação das variáveis mais importantes para o nosso problema, em seguida, selecionaremos o escopo mínimo de variáveis (*feature selection*), por último avaliaremos um conjunto de algoritmos de AM, que serão por fim, comparados. Daremos continuidade ao Workflow adaptado de Hirt, Niklas e Satzger (2017) para desenvolvimento de modelos de AM para problemas de classificação.

## 6.1 INVESTIGAÇÃO DO PROBLEMA

Quadro 10 – Itens identificados na etapa de investigação do problema (Ciclo 1)

Item	Identificação
Partes interessadas ou Stakeholders	<ul style="list-style-type: none"> <li>- Pacientes em tratamento oncológico</li> <li>- Médicos e equipe multidisciplinar de apoio</li> </ul>
Objetivos a serem alcançados	<ul style="list-style-type: none"> <li>- Determinar variáveis relevantes</li> <li>- Determinar conjunto mínimo de variáveis</li> <li>- Escolher algoritmo de AM</li> </ul>
O fenômeno a ser investigado, ou seja, o que será estudado para servir de alternativa para resolução do problema	<ul style="list-style-type: none"> <li>- Os dados fornecidos e pre-processados.</li> </ul>
As causas, mecanismos e razões que dão origem ao problema investigado	<ul style="list-style-type: none"> <li>- Utilizar todos os dados fornecidos em um modelo de AM acarretará em <i>overfitting</i>, levando a baixa performance.</li> <li>- Utilizar um algoritmo inadequado leva à baixa performance.</li> </ul>
As contribuições que a solução proposta pode oferecer	<ul style="list-style-type: none"> <li>- A determinação das variáveis relevantes, conjunto mínimo e melhor algoritmo para esta classe de problemas</li> </ul>

Fonte: O autor.

Para este primeiro ciclo, o problema foi definido nas três questões listadas abaixo:

- Quais variáveis mais importantes para o problema?
- Destas, quais as minimamente necessárias?
- Qual o algoritmo de AM que apresenta melhor desempenho?

Os itens associados ao problema desde ciclo estão resumidos no quadro 10.

## 6.2 DESIGN DA SOLUÇÃO

Para responder a estas questões iniciou-se com a seleção de todas as variáveis importantes no contexto, por meio do algoritmo BORUTA. Em seguida, partindo deste subconjunto

de variáveis selecionado no passo anterior, determinou-se o subconjunto mínimo de variáveis. Por último, construiu-se modelos de AM com algoritmos diferentes e avaliou-se qual mais adequado para ser utilizado.

### 6.2.1 SELEÇÃO DE VARIÁVEIS (FEATURE SELECTION)

Por meio do algoritmo BORUTA, determinou-se que um conjunto de 55 variáveis (do total 349) são relevantes para o problema de classificação em questão sendo, 32 são categóricas e 23 numéricas. A relação de variáveis está listada no Quadro 11. Como realizou-se a codificação *One-hot encoding* cada variável categórica apresentada no quadro 11 é, na verdade, uma alternativa da variável categórica que a originou. Por exemplo, *ATV-Física-Classif\_4.0* representa a resposta [4] da variável *ATV-Física-Classif*. Essa variável tem quatro valores possíveis ([1] Muito ativo, [2] Ativo, [3] Irregularmente Ativo, [4] Sedentário), porém para o problema de classificação em questão, apenas os que se enquadram na categoria Sedentários[4], foram considerados relevantes.

Quadro 11 – Seleção de variáveis. algoritmos BORUTA e RFE.

Boruta	Dimensão	Descrição	Classe	RFE
Altura	Atributo	Altura em metros	Numérica	
CHARLSON	Estado Geral	Indice de Comorbidade de Charlson	Numérica	
PPS	Estado Geral	Escala de performance Paliativa (PPS versão 2)	Numérica	x
Aval-GDS	Estado Mental	Escala de depressão geriátrica	Numérica	x
MINI-MENTAL-Escore	Estado Mental	Avaliação Cognitiva	Numérica	
MINI-MENTAL-Escore30	Estado Mental	Avaliação Cognitiva	Numérica	x
Etil-Ja-Bebeu -Quanto-Tempo	Hábitos Pregressos	Quantidade de anos em que o paciente alega ter bebido ao longo da vida	Numérica	x
ATV-Física-Classif_4.0	Mobilidade	[4]-Sedentário	Categórica	x
Time-Up-Go	Mobilidade	Tempo para completar, em segundos, o teste de mobilidade	Numérica	x
Time-Up-Go- Classi-Mobil_1.0	Mobilidade	[1]-Mobilidade normal	Categórica	x
Time-Up-Go- Classi-Mobil_2.0	Mobilidade	[2]-Anormalidade leve	Categórica	x
Time-Up-Go- Classi-Ris-Qued_1.0	Mobilidade	[1]-Baixo risco de queda	Categórica	
IMC	Nutrição	Índice de massa Corpórea	Numérica	x
MAN-Aval-Est -Nut-Res_1.0	Nutrição	[1]-Risco de desnutrição	Categórica	x
MAN-Aval-Est -Nut-Res_3.0	Nutrição	[3]-Sem problemas de nutrição	Categórica	x
MAN-Aval-Est-Nutric	Nutrição	Escore de nutrição	Numérica	x
MAN-Aval-Global	Nutrição	Mini Avaliação Nutricional - Avaliação Global	Numérica	x
MAN-AvG-G_1.0	Nutrição	[1]-Vive em casa própria	Categórica	x
MAN-AvG-J_2.0	Nutrição	[2]-Faz 3 refeições ao dia	Categórica	x
MAN-AvG-K-1_1.0	Nutrição	[1]-Consume ao menos 1 porção diária de leite/derivados	Categórica	x
MAN-AvG-N_2.0	Nutrição	[2]-Alimenta-se sozinho, sem dificuldade	Categórica	x
MAN-AvG-O_0.0	Nutrição	[0]-Paciente acreditar estar desnutrido	Categórica	x
MAN-AvG-O_2.0	Nutrição	[2]-Não sabe informar se tem problema de nutrição	Categórica	x
MAN-AvG-P_10.0	Nutrição	[10]-Se avalia como tendo boa saúde	Categórica	x
MAN-Escore-Tot	Nutrição	Nota total da avaliação nutricional = Triagem + avaliação global	Numérica	

Quadro 11 continuação da página anterior

Boruta	Dimensão	Descrição	Classe	RFE
MAN-Escore-TR	Nutrição	Nota de triagem da avaliação Nutricional	Numérica	
MAN-TR-A_2.0	Nutrição	[2]-Não houve diminuição recente na ingestão alimentar	Categórica	x
MAN-TR-B_0.0	Nutrição	[0]-Perda de peso superior a 3kg nos últimos meses	Categórica	x
MAN-TR-B_2.0	Nutrição	[2]-Perda de peso entre 1 e 3kg nos últimos meses	Categórica	x
MAN-TR-B_3.0	Nutrição	[3] - Sem perda de peso	Numérica	x
Peso	Nutrição	peso em kg	Categórica	
5-Polifarmacia -Medicamento	Polifarmácia	número de medicamentos de uso continuo, receitados ou não.	Numérica	x
QLQ30-01_1	Qual. de vida	[1]-Não tem dificuldade em fazer esforços físicos	Categórica	x
QLQ30-02_1	Qual. de vida	[1]-Não tem dificuldade em fazer caminhadas longas	Categórica	x
QLQ30-03_1	Qual. de vida	[1]-Não tem dificuldade em caminhadas curtas	Categórica	x
QLQ30-04_1	Qual. de vida	[1]-Não fica em cadeira/cama o dia todo	Categórica	x
QLQ30-07_1	Qual. de vida	[1]-Não tem sido difícil para se divertir	Categórica	
QLQ30-09_1	Qual. de vida	[1]-Não teve dor na última semana	Categórica	x
QLQ30-09_2	Qual. de vida	[2]-Sentiu um pouco de dor na última semana	Categórica	x
QLQ30-10_1	Qual. de vida	[1]-Não precisou repousar mais que costume na última semana	Categórica	x
QLQ30-12_1	Qual. de vida	[1]-Não se sentiu fraco na última semana	Categórica	
QLQ30-13_1	Qual. de vida	[1]-Não teve falta de apetite na última semana	Categórica	x
QLQ30-18_1	Qual. de vida	[1]-Não esteve cansado na última semana	Categórica	x
QLQ30-19_1	Qual. de vida	[1]-Na última semana, não sentiu dor a ponto de atrapalhar a rotina	Categórica	x
QLQ30-28_1	Qual. de vida	[1]-Na última semana, o tratamento não trouxe dificuldades financeiras	Categórica	x
QLQ30-29_6	Qual. de vida	[6]-Se avalia como tendo uma boa saúde(6/7)	Categórica	x
QLQ30-29_7	Qual. de vida	[7]-Se avalia como tendo ótima saúde	Categórica	x
Tabag-Q.02.b	Qual. de vida	Qual idade tinha quando parou de fumar(em anos)	Numérica	
Contag-Leucocitos -Result	Result. de exames	Contagem de Leucócitos	Numérica	x
Creatinina-Result	Result. de exames	Nível de creatinina	Numérica	x
Granulocitos-Result	Result. de exames	Contagem de Granulócitos	Numérica	x
Hemoglobina-Result	Result. de exames	Contagem Hemoglobina	Numérica	x
Plaquetas-Result	Result. de exames	Contagem de Plaquetas	Numérica	
Q.05	Sócio econômicas	Renda doméstica	Numérica	
CID10_61.0	Tipo de Câncer	[61]-Câncer de próstata	Categórica	x

Fonte: O autor

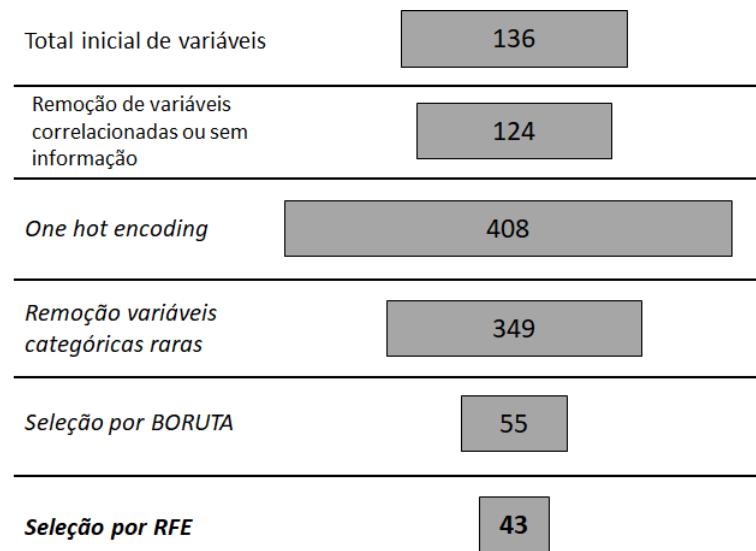
As variáveis selecionadas são úteis para o entendimento do problema, no entanto ainda não formam o conjunto mínimo. Para determiná-lo, utilizou-se o algoritmo RFE (*Recursive Feature Elimination*) que, conforme demonstrado por Cueto-López et al. (2019), quando utilizado em conjunto com os algoritmos Máquinas Vetoriais de Suporte (SVM) ou Regressão Logística (LR) como avaliadores, mostra-se adequado para realizar a seleção de variáveis para um problema de classificação de risco de pacientes com câncer colo-retal. Utilizou-se o RFE para avaliar exaustivamente, de 1 a 55 variáveis, qual o sub-conjunto mínimo, onde o resultado obtido foi uma seleção de 43 delas, resumidas no Quadro 12. Neste Quadro, as variáveis encontram-se separadas por dimensão. A quantidade de variáveis, do início ao fim do processo, está resumida na Figura 15.

Quadro 12 – Resumo do número de variáveis utilizadas por dimensão de avaliação.

Dimensão	Total de variáveis relevantes (Boruta)	Total de variáveis necessárias (RFE)
Nutrição	19	16
Qualidade de vida	16	13
Resultados clínicos	5	4
Mobilidade	5	4
Estado Geral	2	1
Estado Mental	3	2
Tipo de Câncer	1	1
Polifarmácia	1	1
Sócio Econômicas	1	0
Atributo físico	1	0
Hábitos pregressos	1	1
Total	55	43

Fonte: O autor.

Figura 15 – Seleção de variáveis do início ao fim do processo.



Fonte: O autor.

Através de uma análise quantitativa em cada dimensão verifica-se que:

- **Nutrição:** As 16 variáveis desta dimensão, selecionadas pelo RFE, pertencem ao questionário de *Mini Avaliação Nutricional - MAN*, de acordo com Guigoz (2006), o MAN (ou MNA - *Mini Nutritional Assessment*) é realizado em duas etapas. Na primeira etapa, chamada de MNA-SF (*Short Form*), são realizadas 6 perguntas (A, B, C, D, E e F) e calculado um escore de triagem. Caso o paciente some 12 ou mais pontos, já é considerado com nutrição normal e é desnecessário realizar o restante das perguntas. Caso a pontuação seja inferior aos 12 pontos, as demais 12 perguntas são realizadas (perguntas de G a R). O algoritmo RFE selecionou: Escore de nutrição (*MAN-Aval-Est-Nutric*), nota global da avaliação (*MAN-Aval-Global*), IMC (que não faz parte do formulário MAN, mas do mesmo contexto) além de outras 13 respostas específicas do formulário como, por exemplo, se o paciente respondeu a questão J especificamente com o valor [2] - *realiza três refeições por dia*. O formulário da MAN encontra-se na página 15 do ANEXO I.
- **Qualidade de vida:** Incluídas 13 variáveis categóricas relacionadas a respostas específicas do questionário EORTC-QLQ30 (European Organization for Research and Treatment of Cancer - Quality of Life Questionnaire), como: [1]-Não tem dificuldade em fazer esforços físicos, [1]-Não teve falta de apetite na última semana entre outras listadas no Quadro 11.
- **Resultados clínicos:** Resultados de 4 exames: Contagem de Hemoglobinas, Leucócitos, Creatinina, Granulócitos.
- **Mobilidade:** 4 variáveis selecionadas ,sendo uma numérica: resultado do teste *Timed UP and GO* em segundos, se o paciente foi classificado como [4]-Sedentário, avaliação da mobilidade ([1]-Mobilidade normal ou [2]-Anormalidade leve) e risco de queda([1]-Baixo risco de queda).
- **Estado Geral:** Valor numérico da Escala de performance Paliativa (PPS versão 2). A Escala PPS varia de 100%, que significa máxima atividade funcional, até 0%, indicando morte.
- **Estado Mental:** Foram incluídas o resultado da avaliação de depressão geriátrica (Aval-GDS) e o resultado do teste cognitivo *MINI-MENTAL-Escore30*.
- **Tipo de Câncer:** De todos os cânceres presentes na base de dados, apenas câncer de próstata foi selecionado para fazer parte do conjunto de variáveis capaz de predizer risco de óbito em 6 meses. O câncer de próstata, de acordo com Oliveira et al. (2016), é o de maior incidência em homens e também o que mais mata.

- **Polifarmácia:** Número de medicamentos, receitados ou não, que o paciente faz uso regular.
- **Sócio Econômicas:** A Renda doméstica foi selecionada pelo Boruta como sendo um fator relevante, porém considerada não essencial pela análise realizada com o algoritmo RFE.
- **Atributo Físico:** Altura em metros. Foi descartada pelo RFE pelo fato da informação estar presente em outra variável (IMC)
- **Hábitos Progressos:** Quantidade de anos em que o paciente alega ter bebido ao longo da vida, foi considerado uma variável importante para predição de óbito em 6 meses.

Desta forma, respondeu-se a primeira parte do problema definido para este ciclo. Conforme exposto acima, 55 variáveis foram consideradas relevantes para o problema, sendo 43 deles consideradas necessárias.

### 6.2.2 SELEÇÃO DE ALGORITMOS

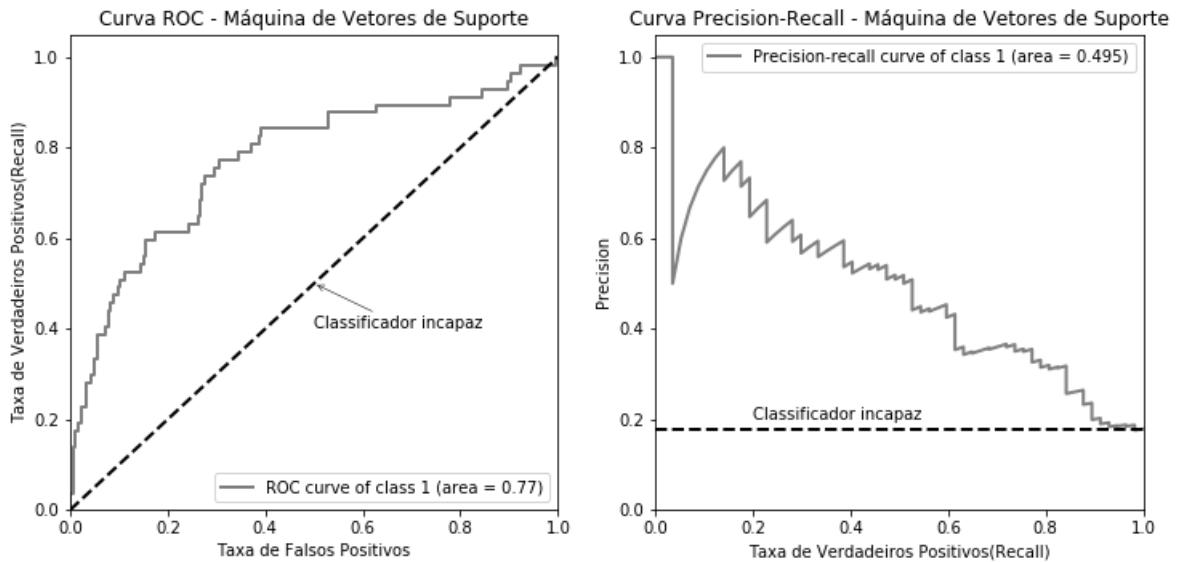
A segunda parte do problema definido neste ciclo foi: Qual melhor algoritmo de AM para o problema em questão?

Para responder a esta pergunta, realizamos um *experimento* onde avaliamos a performance dos 7 algoritmos listados abaixo:

- Regressão Logística
- Análise Discriminante Linear
- K Vizinhos Próximos
- Naive Bayes
- Máquina de Vetores de Suporte
- Perceptron Multicamada
- XGboost

Considerou-se, para a construção dos modelos, as 43 variáveis selecionadas na seção anterior como necessárias para o nosso problema. Utilizou-se a base de Treinamento/Validação, com validação cruzada com  $k = 10$ . A base de Teste foi utilizada para calcular o desempenho, em conjunto com as métricas definidas anteriormente: *AUC-ROC*, *Recall* e *F1-score*,

Figura 16 – Curvas ROC e *Precision-Recall* para o algoritmo Máquina de Vetores de Suporte (SVM).



Fonte: O autor

como métricas informativas e *AUC-Precision-Recall* como determinante para de decisão do melhor algoritmo.

Os algoritmos selecionados não foram otimizados (estamos utilizando nesta avaliação os valores padrão do Scikit-Learn), a menos do Perceptron Multicamadas que necessita da informação do design da rede neural. Segundo princípios de design e heurísticas propostas por Walczak e Cerpa (1999), utilizamos uma rede neural com 2x a quantidade de variáveis (43) e um número de camadas internas igual ao número de passos que o especialista (médico) utilizaria para resolver o problema. Consideramos cada domínio do Quadro 12, menos os domínios com apenas uma variável, como um passo. Dessa forma a rede neural será composta por 5 camadas internas de 86 neurônios cada (86,86,86,86,86).

Conforme resumido na Tabela 5, o algoritmo com melhor desempenho foi Análise Discriminante Linear (LDA). As Figuras 16 a 22 estão dispostas as curvas ROC, *Precision-Recall* e matriz de confusão para cada um dos 7 modelos ajustados utilizando os algoritmos listados.

Tabela 5 – Resumo de desempenho dos algorítmicos avaliados no ciclo 1 de DSR.

algoritmo	Precisão	Recall	Taxa Falsos Positivos	AUC ROC	AUC Precision-Recall	F1 score
Análise Discriminante Linear	0,547	0,614	0,110	<b>0,840</b>	<b>0,632</b>	<b>0,579</b>
Regressão Logística	0,374	<b>0,754</b>	0,274	0,823	0,611	0,500

**Tabela 5** continuada da página anterior

algoritmo	Precisão	Recall	Taxa Falsos Positivos	AUC ROC	AUC Precision-Recall	F1 score
Perceptron Multicamada	<b>0,583</b>	0,491	<b>0,076</b>	0,815	0,570	0,533
Naive Bayes	0,436	0,421	0,118	0,775	0,529	0,428
XGboost	0,327	0,649	0,289	0,761	0,528	0,435
Máquina de Vetores de Suporte	0,500	0,526	0,114	0,771	0,495	0,513
K Vizinhos Próximos	0,488	0,702	0,160	0,826	0,491	0,576
<b>Classificador de Referência</b>	0,333	0,404	0,175	0,683	0,372	0,365

Fonte: O autor.

### 6.3 VALIDAÇÃO DA SOLUÇÃO

Retornando às questões referentes ao problema deste ciclo:

- Quais variáveis mais importantes para o problema?
- Destas, quais as minimamente necessárias?
- Qual o algoritmo de AM que apresenta melhor desempenho?

Determinou-se o conjunto de 55 variáveis relevantes ao problema em questão, sendo 43 minimamente necessárias ao modelo de AM conforme Quadro 12.

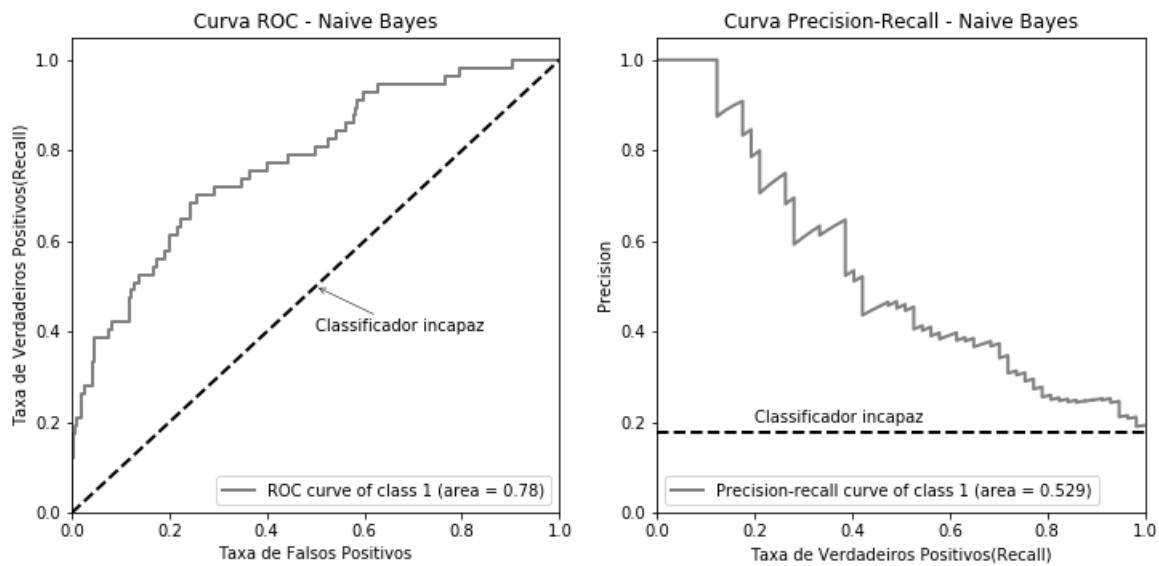
De acordo com a tabela 5, o algoritmo Análise Discriminante Linear (Figura 22) obteve a maior *AUC Precision-Recall* (0,632) e a maior AUC ROC (0,840) quando analisado com dados não vistos (dados de teste).

Verifica-se que, em comparação com o classificador básico criado no final do capítulo 5 utilizando todas as variáveis de entrada, houve um melhora significativa em todos indicadores.

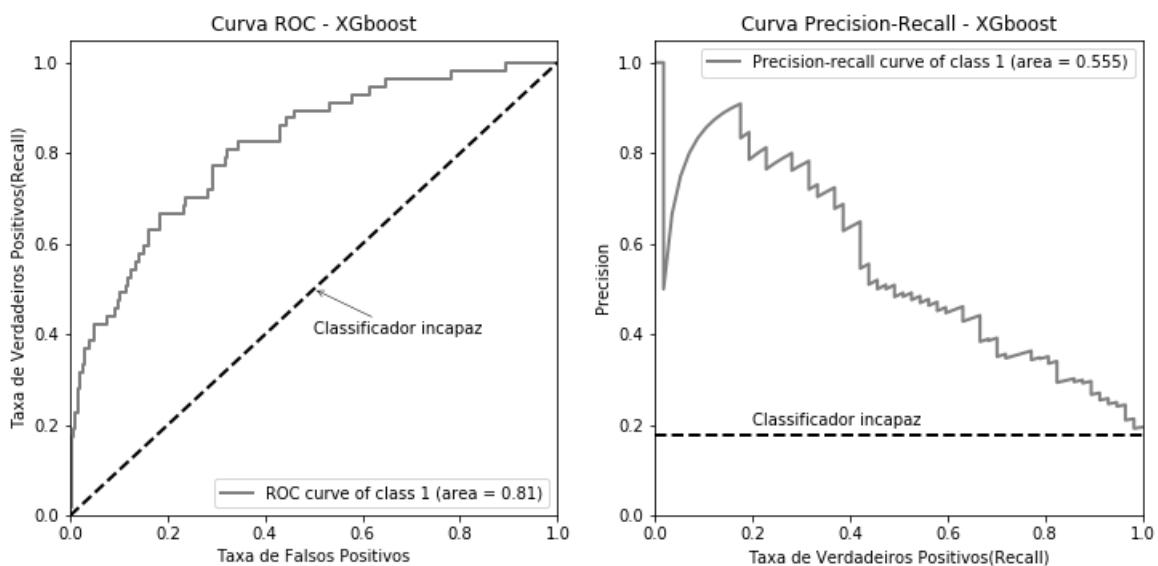
Dado um classificador operando no ponto 1, a matriz de confusão para os dados de teste é a mostrada na Figura 22.

O classificador de referência identificou corretamente 23 pacientes com risco de óbito e não detectou outros 34. O novo classificador, utilizando o algoritmo de análise discriminante linear e um conjunto menor de features, detectou 35 pacientes com risco de óbito e não detectou outros 22. Uma melhora significativa.

Há de se pontuar que os demais algoritmos poderiam obter desempenhos melhores, caso otimizados. O K-NN foi utilizado com valor padrão da implementação do Sci-kit

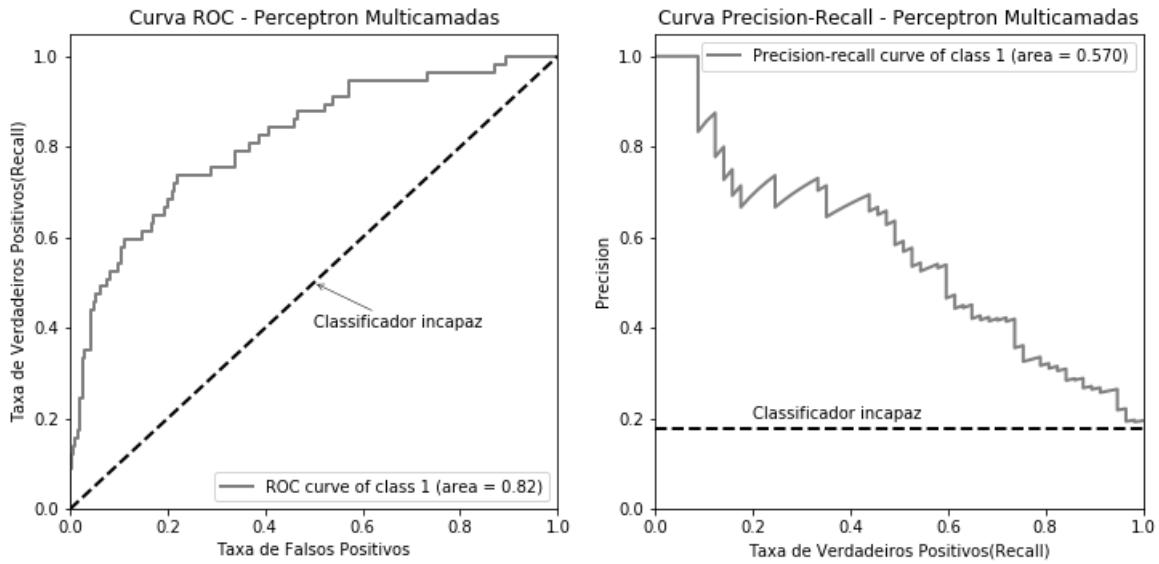
Figura 17 – Curvas ROC e *Precision-Recall* para o algoritmo Naive Bayes (NB).

Fonte: O autor

Figura 18 – Curvas ROC e *Precision-Recall* para o algoritmo XGBoost.

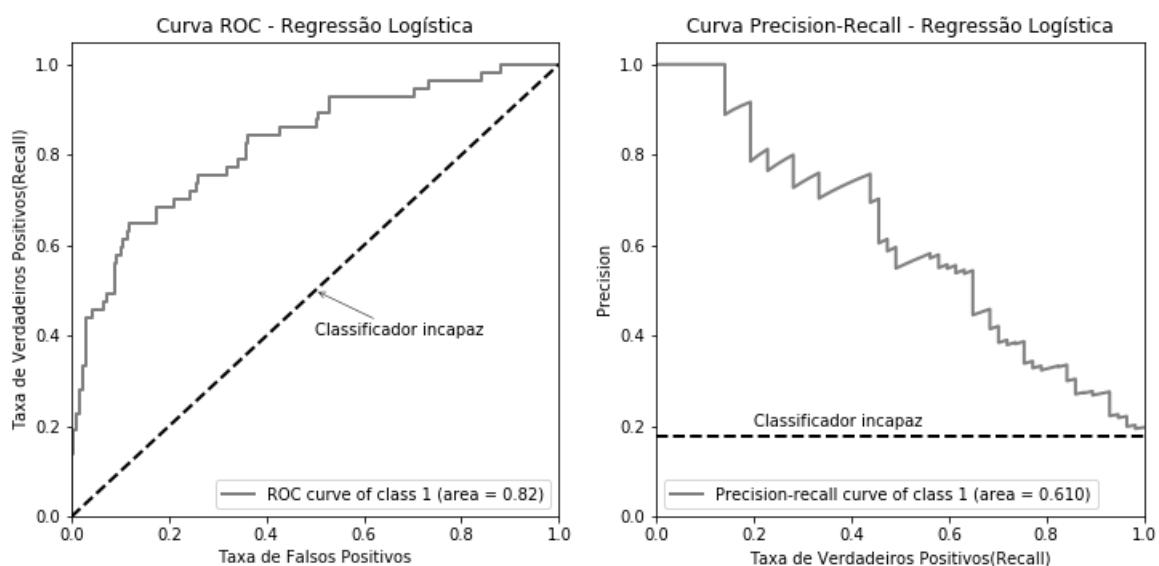
Fonte: O autor

Figura 19 – Curvas ROC e *Precision-Recall* para o algoritmo - Perceptron Multicamadas (MLP).



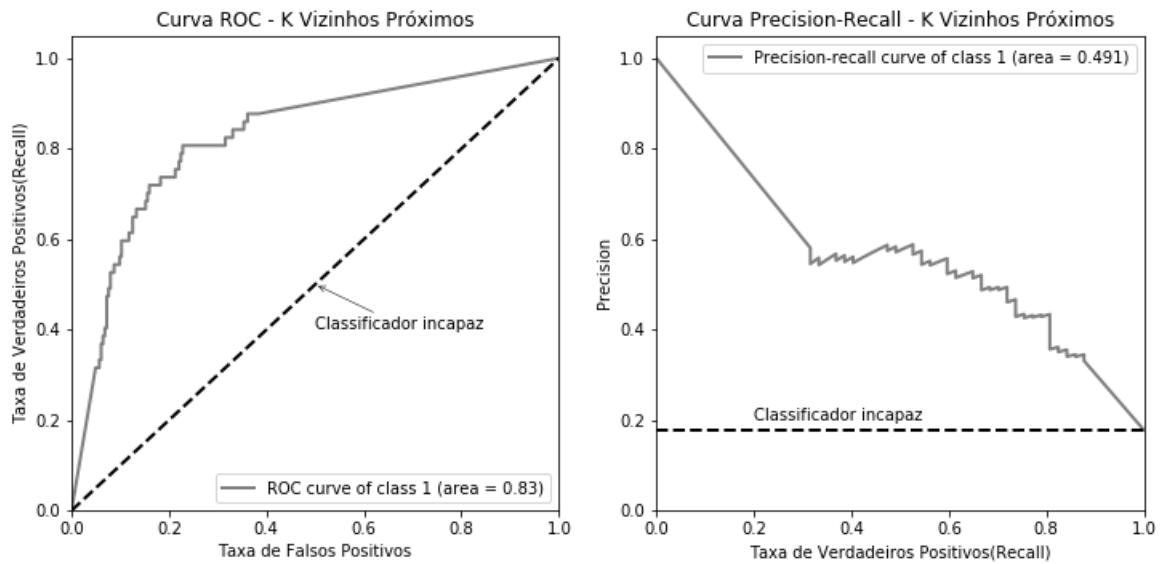
Fonte: O autor

Figura 20 – Curvas ROC e *Precision-Recall* para o algoritmo Regressão Logística (LR).



Fonte: O autor

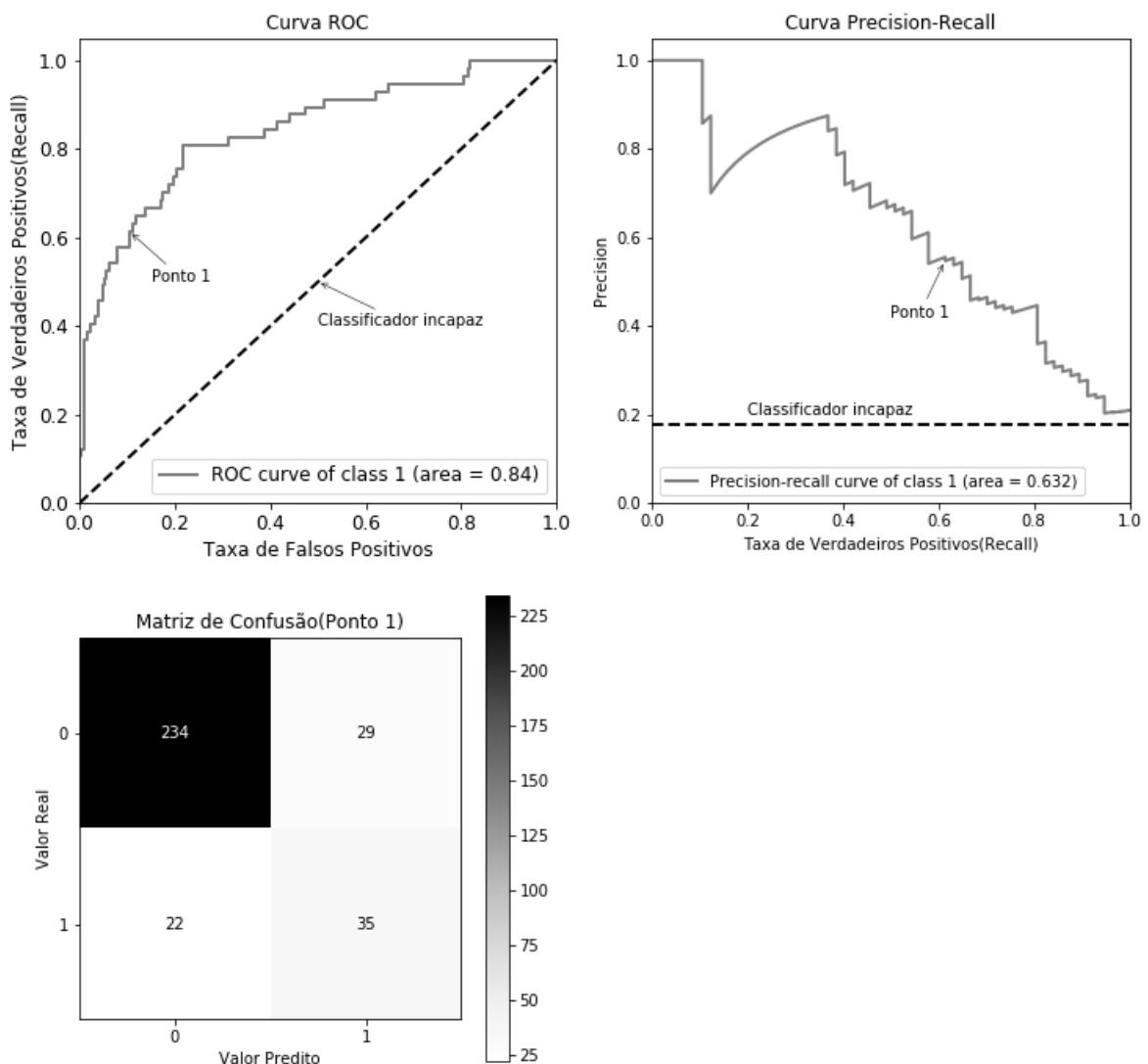
Figura 21 – Curvas ROC e *Precision-Recall* para o algoritmo K Vizinhos Próximos (KNN).



Fonte: O autor

Learn (número de vizinhos igual a 5), o Perceptron Multicamadas tem seu desempenho determinado pelo design da rede, da mesma forma ocorre com os outros algoritmos utilizados. No próximo ciclo realizou-se a otimização de performance de cada um deles.

Figura 22 – Curvas ROC, *Precision-Recall* e matriz de confusão (exemplo) para o algoritmo Análise Discriminante Linear (LDA).



Fonte: O autor

## 7 CICLO 2 - OTIMIZAÇÃO

Este capítulo tem como objetivo apresentar o segundo ciclo de design, conforme identificado no capítulo 4. Dando continuidade ao *workflow* (Figura 12) adaptado de Hirt, Niklas e Satzger (2017) para desenvolvimento de modelos de AM para problemas de classificação iremos, neste ciclo, realizar a otimização dos modelos de AM do ciclo anterior, conforme descrita na etapa *Estimar Erro do modelo*.

### 7.1 INVESTIGAÇÃO DO PROBLEMA

Quadro 13 – Itens identificados na etapa de investigação do problema (Ciclo 2)

Item	Identificação
Partes interessadas ou Stakeholders	- Pacientes em tratamento oncológico - Médicos e equipe multidisciplinar de apoio
Objetivos a serem alcançados	- Obter um modelo de AM desenvolvido para melhor identificação de Pacientes de Alto risco.
O fenômeno a ser investigado, ou seja, o que será estudado para servir de alternativa para resolução do problema	Os modelos desenvolvidos no ciclo anterior.
As causas, mecanismos e razões que dão origem ao problema investigado	Existem parâmetros de ajuste nos modelos de AM que podem melhorar o desempenho dos mesmos
As contribuições que a solução proposta pode oferecer	Uma melhor classificação dos pacientes de alto risco de óbito em 6 meses.

Fonte: O autor.

No ciclo 1 foi possível desenvolver modelos com desempenho melhor que o classificador de referência. No entanto ainda é possível melhorar a performance, por meio de ajustes em cada um dos algoritmos utilizados. O ciclo 2 de design teve início com o estudo dos algoritmos utilizados e quais parâmetros de ajuste relevantes para cada um deles.

Para este segundo ciclo, o problema definido foi:

- Como otimizar os algoritmos para máximo desempenho considerando a métrica *AUC Precision-Recall?*

Os itens associados ao problema desde ciclo estão resumidos no Quadro 13.

### 7.2 DESIGN DA SOLUÇÃO

Para responder a esta questão realizou-se uma busca em grade (*gridsearch*) em um espaço de parâmetros definido para cada algoritmo. Por fim, o modelo de melhor desempenho será escolhido como o melhor adequado ao problema em estudo.

### 7.2.1 ANÁLISE DE DISCRIMINANTE LINEAR

O Algoritmo de Análise de Discriminante Linear (LDA) tem uma solução de forma fechada e, portanto, não tem hiperparâmetros para serem ajustados. Porém existem algumas opções que podem ser testadas e que foram brevemente explicados na seção 2.2.7. No Quadro 14 estão descritos os valores que foram utilizados na busca em grade.

Quadro 14 – Espaço de busca para o LDA.

```
priors_values = [(0.1,0.9),(0.15,0.85),(0.2,0.8),(0.25,0.75),(0.3,0.7),(0.35,0.65),
(0.4,0.6),(0.45,0.55),(0.5,0.5)]
solver_values = ['svd', 'lsqr', 'eigen']
shrin_values = ['none', 'auto', 0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1]
```

Fonte: O autor

Finalizou-se, portanto, com **3 x 13 x 9 = 351** combinações diferentes para teste com validação cruzada,  $k = 10$ , resultando em 3510 iterações. A melhor solução utilizou a proporção de balanceamento de classes de (0.3, 0,7), o fator *shrinkage* = 0 e o algoritmo de otimização *lsqr*, conforme apresentado no trecho da execução do código no Quadro 15.

Quadro 15 – Busca exaustiva realizada em 351 combinações diferentes de 3 parâmetros, além de validação cruzada.

```
StartFragmentFitting 10 folds for each of 351 candidates, totalling 3510 fits
[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent workers.
Best: 0.965176 using {'priors': (0.3, 0.7), 'shrinkage': 0, 'solver': 'lsqr'}
nan (nan) with: {'priors': (0.1, 0.9), 'shrinkage': 'none', 'solver': 'svd'}
nan (nan) with: {'priors': (0.1, 0.9), 'shrinkage': 'none', 'solver': 'lsqr'}
nan (nan) with: {'priors': (0.1, 0.9), 'shrinkage': 'none', 'solver': 'eigen'}

.
.
.
3510 iterações
.

nan (nan) with: {'priors': (0.5, 0.5), 'shrinkage': 1, 'solver': 'svd'}
0.900045 (0.015426) with: {'priors': (0.5, 0.5), 'shrinkage': 1, 'solver': 'lsqr'}
0.900045 (0.015426) with: {'priors': (0.5, 0.5), 'shrinkage': 1, 'solver': 'eigen'}
[Parallel(n_jobs=1)]: Done 3510 out of 3510 | elapsed: 1.0min finished
EndFragment
```

Fonte: O autor.

### 7.2.2 REGRESSÃO LOGÍSTICA

Os parâmetros para a Regressão Logística foram brevemente explicados na seção 2.2.7. No Quadro 16 estão descritos os valores que foram utilizados na busca em grade.

Quadro 16 – Espaço de busca para a Regressão Logística.

```
c_values = [0.1, 0.3, 0.5, 0.7, 0.9, 1.0, 1.3, 1.5, 1.7, 2.0]
solver_values = ['newton-cg', 'lbfgs', 'liblinear', 'sag', 'saga']
```

Fonte: O autor.

Finalizou-se, portanto, com  $5 \times 10 = 50$  diferentes combinações para teste com validação cruzada,  $k = 10$ , sendo portanto 500 iterações. A melhor solução utilizou o parâmetro  $C=0.3$  e o algoritmo de otimização *liblinear*, conforme apresentado no trecho da execução do código no Quadro 17.

Quadro 17 – Busca exaustiva realizada em 50 combinações diferentes de 2 parâmetros, além de validação cruzada.

```
Fitting 10 folds for each of 50 candidates, totalling 500 fits
[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent workers.
Best: 0.916745 using {'C': 0.3, 'solver': 'liblinear'}
0.913931 (0.021958) with: {'C': 0.1, 'solver': 'newton-cg'}
0.913931 (0.021958) with: {'C': 0.1, 'solver': 'lbfgs'}
0.911995 (0.021611) with: {'C': 0.1, 'solver': 'liblinear'}
0.913931 (0.021958) with: {'C': 0.1, 'solver': 'sag'}
.
.
.
500 iterações
.
.
.
0.914186 (0.018185) with: {'C': 2.0, 'solver': 'liblinear'}
0.915584 (0.019443) with: {'C': 2.0, 'solver': 'sag'}
0.915003 (0.020488) with: {'C': 2.0, 'solver': 'saga'}
[Parallel(n_jobs=1)]: Done 500 out of 500 | elapsed: 49.0s finished
```

Fonte: O autor.

### 7.2.3 NAIVE BAYES

Os parâmetros para o algoritmo Naive Bayes foram brevemente explicados na seção 2.2.7. No Quadro 18 estão descritos os valores que foram utilizados na busca em grade.

Quadro 18 – Espaço de busca para o algoritmo Naive Bayes

```
priors_values = [(0.1,0.9),(0.15,0.85),(0.2,0.8),(0.25,0.75),
(0.3,0.7),(0.35,0.65),(0.4,0.6),(0.45,0.55),(0.5,0.5)]
```

Fonte: O autor.

Finalizamos, portanto, com **9** combinações diferentes para teste com validação cruzada,  $k = 10$ , sendo portanto 90 iterações. A melhor solução utilizou o parâmetro de prioridade

igual a  $(0.1, 0.9)$ , conforme apresentado no trecho da execução do código no Quadro 19.

Quadro 19 – Busca exaustiva realizada em 9 combinações diferentes de 1 parâmetro, além de validação cruzada.

```
Fitting 10 folds for each of 9 candidates, totalling 90 fits
[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent workers.
Best: 0.891418 using {'priors': (0.1, 0.9)}
0.891418 (0.019600) with: {'priors': (0.1, 0.9)}
0.890251 (0.019627) with: {'priors': (0.15, 0.85)}
0.888231 (0.019222) with: {'priors': (0.2, 0.8)}
0.887974 (0.021744) with: {'priors': (0.25, 0.75)}
0.886485 (0.022712) with: {'priors': (0.3, 0.7)}
0.885902 (0.022185) with: {'priors': (0.35, 0.65)}
0.884779 (0.021172) with: {'priors': (0.4, 0.6)}
0.883752 (0.021283) with: {'priors': (0.45, 0.55)}
0.884128 (0.021436) with: {'priors': (0.5, 0.5)}
[Parallel(n_jobs=1)]: Done, 90 out of 90 | elapsed: 0.7s finished
```

Fonte: O autor.

#### 7.2.4 MÁQUINA DE VETORES DE SUPORTE

Os parâmetros para o algoritmo Máquina de Vetores de Suporte (SVM) foram brevemente explicados na seção 2.2.7. No Quadro 20 estão descritos os valores que foram utilizados na busca em grade.

Quadro 20 – Espaço de busca para o algoritmo Máquina de Vetores de Suporte (SVM)

```
c_values = [0.1, 0.3, 0.5, 0.7, 0.9, 1.0, 1.3, 1.5, 1.7, 2.0]
kernel_values = ['linear', 'poly', 'rbf', 'sigmoid']
```

Fonte: O autor.

Finalizou-se, portanto, com  $10 \times 4 = 40$  combinações diferentes para teste com validação cruzada,  $k = 10$ , sendo portanto 400 iterações. A melhor solução utilizou o parâmetro  $C=1.7$  e  $kernel='rbf'$ , conforme apresentado no trecho da execução do código no Quadro 21.

Quadro 21 – Busca exaustiva realizada em 40 combinações diferentes de 2 parâmetros, além de validação cruzada.

```
Fitting 10 folds for each of 40 candidates, totalling 400 fits
[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent workers.
[Parallel(n_jobs=1)]: Done 400 out of 400 | elapsed: 6.0min finished
Best: 0.925691 using {'C': 1.7, 'kernel': 'rbf'}
0.919599 (0.016599) with: {'C': 0.1, 'kernel': 'linear'}
0.868677 (0.021223) with: {'C': 0.1, 'kernel': 'poly'}
0.905076 (0.016073) with: {'C': 0.1, 'kernel': 'rbf'}
.
. 400 iterações
```

Fonte: O autor.

### 7.2.5 K VIZINHOS PRÓXIMOS

Os parâmetros para o algoritmo KNN foram brevemente explicados na seção 2.2.7. No Quadro 22 estão descritos os valores que foram utilizados na busca em grade.

Quadro 22 – Espaço de busca para o algoritmo KNN.

```
n_values = [4,5,6,7,8,9,10,11,12,13,14,15]
weights_values = ['uniform', 'distance']
algorithm_values = ['ball tree', 'kd tree', 'brute']
```

Fonte: O autor.

Finalizou-se, portanto, com  $12 \times 2 \times 3 = 72$  combinações diferentes para teste com validação cruzada,  $k = 10$ , sendo portanto 720 iterações. A melhor solução utilizou o algoritmo *ball\_tree*,  $n\_neighbors=4$  e  $weight='uniform'$ , conforme apresentado no trecho da execução do código no Quadro 23.

Quadro 23 – Busca exaustiva realizada em 72 combinações diferentes de 3 parâmetros, além de validação cruzada.

```
Fitting 10 folds for each of 72 candidates, totalling 720 fits
[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent workers.
Best: 0.906501 using {'algorithm': 'ball_tree', 'n_neighbors': 4, 'weights': 'uniform'}
0.906501 (0.009984) with: {'algorithm': 'ball_tree', 'n_neighbors': 4, 'weights': 'uniform'}
0.904246 (0.014013) with: {'algorithm': 'ball_tree', 'n_neighbors': 4, 'weights': 'distance'}
0.895110 (0.009948) with: {'algorithm': 'ball_tree', 'n_neighbors': 5, 'weights': 'uniform'}
```

Fonte: O autor.

### 7.2.6 XGBOOST

Os parâmetros para o algoritmo XGBoost foram brevemente explicados na seção 2.2.7. No Quadro 24 estão descritos os valores que foram utilizados na busca em grade.

Quadro 24 – Espaço de busca para o algoritmo XGBoost.

scale_pos_weight_values = [0.1, 0.3, 0.5, 0.9, 1]
learning_rate_values = [0.01, 0.05, 0.07, 0.1]
max_depth_values = [3, 5, 9, 11, 15, 19, 21, 25, 27, 31, 35, 37, 41]
subsample_values = [0.8, 0.9, 1]
colsample_bytree_values = [0.3, 0.5, 0.7, 0.8]
gamma_values = [0, 1, 5]

Fonte: O autor.

Finalizamos, portanto, com **5 x 4 x 13 x 3 x 4 x 2 = 9360** combinações diferentes para teste com validação cruzada,  $k = 3$ , sendo portanto 28080 iterações. A melhor solução utilizou os parâmetros  $\text{colsample\_bytree}=0.3$ ,  $\text{gamma}=0$ ,  $\text{learning\_rate}=0.07$ ,  $\text{max\_depth}=15$ ,  $\text{scale\_pos\_weight}=0.5$  e  $\text{subsample}=0.9$ , conforme apresentado no trecho da execução do código no Quadro 25. Esta busca em particular levou cerca de 3h em um notebook atual (Core i5 vPro 64bits, 8ª Geração, 8Gb DDRAM).

Quadro 25 – Busca exaustiva realizada em 9360 combinações.

Fitting 3 folds for each of 9360 candidates, totalling 28080 fits
[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent workers.
[Parallel(n_jobs=1)]: Done 28080 out of 28080   elapsed: 171.0min finished
Best: 0.932465 using {'colsample_bytree': 0.3, ..., 'subsample': 0.9}
0.316991 (0.025195) with: {'colsample_bytree': 0.3, ..., 'subsample': 0.8}
0.326151 (0.016293) with: {'colsample_bytree': 0.3, ..., 'subsample': 0.9}

Fonte: O autor.

### 7.2.7 PERCEPTRON MULTICAMADA

Os parâmetros para o algoritmo Perceptron Multicamada foram brevemente explicados na seção 2.2.7. No Quadro 26 estão descritos os valores que foram utilizados na busca em grade.

Quadro 26 – Espaço de busca para o algoritmo MLP.

```

solver_values = ['lbfgs', 'sgd', 'adam']
alpha_values = [0.0001, 0.05, 0.01, 0.5]
activation_values = ['identity', 'logistic', 'tanh', 'relu']
hidden_layer_values = [(86, 86, 86, 86), (43, 43, 43, 43),
(172, 172, 172, 172),
(86, 86, 86, 86, 86, 86, 86, 86, 86)]
learning_values = ['constant', 'invscaling', 'adaptive']

```

Fonte: O autor.

Finalizamos, portanto, com **3 x 4 x 4 x 6 x 3 = 864** combinações diferentes para teste com validação cruzada,  $k = 10$ , sendo portanto 8640 iterações. A melhor solução utilizou os parâmetros  $activation = 'relu'$ ,  $alpha=0.001$ ,  $hidden\_layers = (86, 86, 86, 86)$ ,  $learning\_rate = 'adaptative'$  e  $solver = 'sgd'$ , conforme apresentado no trecho da execução do código no Quadro 27.

Quadro 27 – Busca exaustiva realizada com 864 combinações de 5 parâmetros diferentes, além de validação cruzada.

```

Fitting 3 folds for each of 288 candidates, totalling 864 fits
[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent workers.
[Parallel(n_jobs=1)]: Done 864 out of 864 | elapsed: 49.8min finished
Best: 0.921254 using {...hidden_layer_sizes': (86, 86, 86, 86, 86), ... 'solver': 'sgd'}
0.443226 (0.313410) with: {.... 'hidden_layer_sizes': (86, 86, 86, 86, 86),... 'solver': 'lbfgs'}
0.666659 (0.002760) with: {...'hidden_layer_sizes': (86, 86, 86, 86, 86), ... 'solver': 'sgd'}
0.912184 (0.006748) with: {...'hidden_layer_sizes': (86, 86, 86, 86, 86),... 'solver': 'adam'}
.
.
.
```

Fonte: O autor.

### 7.3 VALIDAÇÃO DA SOLUÇÃO

Retornando ao problema identificado para este ciclo:

- Como otimizar os algoritmos para máximo desempenho considerando a métrica *AUC Precision-Recall*?

Conforme pode ser verificado da Tabela 6, o modelo utilizando **Máquina de Vetores de Suporte - SVM** obteve uma ligeira melhora de desempenho na métrica *AUC Precision-Recall* (0,660) em relação ao melhor modelo do anterior (ciclo 1) obtido pelo algoritmo Análise Discriminante Linear (0,632).

Tabela 6 – Resumo de desempenho dos algoritmos avaliados no ciclo 2 de DSR. Após otimização realizada, houve melhora no desempenho em dados não vistos (dados de teste) em todos eles.

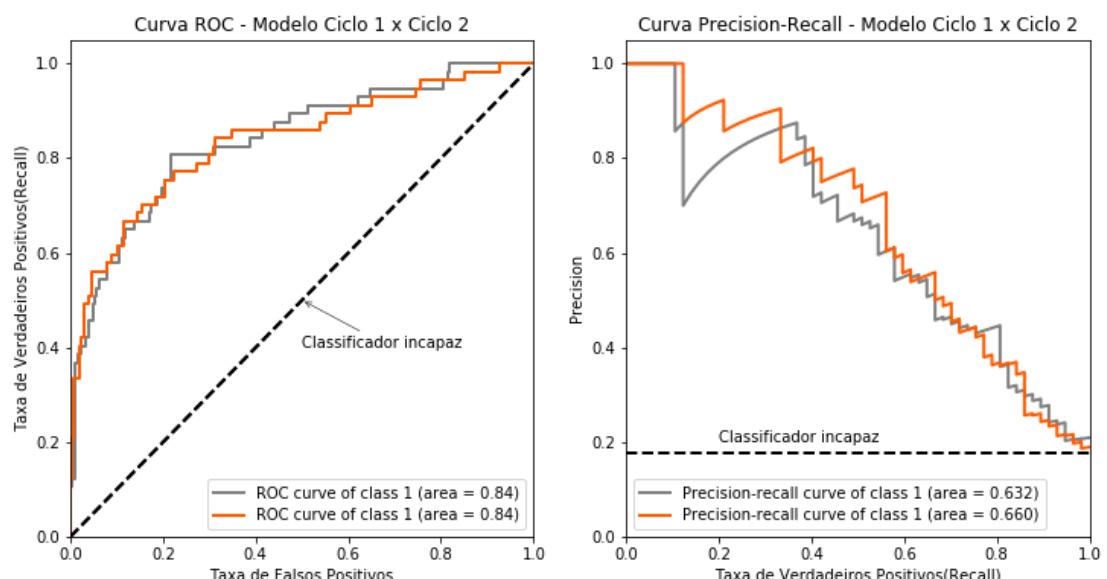
Algoritmo	Precisão	Recall	Taxa Falsos Positivos	AUC ROC	AUC PR	F1
Máquina de Vetores de Suporte	<b>0,653</b>	0,561	<b>0,065</b>	0,836	<b>0,660</b>	<b>0,604</b>
Regressão Logística	0,574	0,614	0,099	0,840	0,626	0,593
Análise Discriminante Linear	0,384	<b>0,842</b>	0,293	<b>0,843</b>	0,619	0,527
XGBoost	0,435	0,684	0,179	0,828	0,585	0,532
Perceptron Multicamadas	0,520	0,456	0,091	0,789	0,541	0,486
Naive Bayes	0,458	0,474	0,122	0,775	0,529	0,466
K Vizinhos Próximos	0,507	0,649	0,137	0,807	0,436	0,569
Classificador de Referência	0,333	0,404	0,175	0,683	0,372	0,365

Fonte: O autor.

A Figura 23 mostra as curvas ROC e *Precision-Recall* para os dois modelos gerados (ciclos 1 e 2), onde é possível observar que o classificador do ciclo 2 é ligeiramente melhor que o gerado no ciclo 1.

Dessa forma encerra-se a etapa de validação do o segundo ciclo de DSR, conseguindo melhorar a performance do modelo gerado no ciclo 1 quando avaliado sob a métrica AUC *Precision-Recall*. O modelo identificado no ciclo 2, alcançou os melhores valores em 4 dos 6 indicadores considerados, inclusive na AUC *Precision-Recall*, sendo portanto, nossa escolha para o modelo de classificação. Observa-se que a melhora não foi expressiva da Figura 23 não se justificando a execução de um terceiro ciclo de design.

Figura 23 – Curvas ROC, *Precision-Recall* para os modelos do ciclo 1 (cinza) e ciclo 2 (vermelho).



Fonte: O autor

## 8 CONCLUSÕES

O objetivo desse capítulo é apresentar as considerações finais e contribuições, assim como resultados alcançados, ameaças à validade, limitações e trabalhos futuros decorrentes dessa pesquisa.

### 8.1 CONSIDERAÇÕES FINAIS E CONTRIBUIÇÕES

Na primeira fase deste trabalho realizou-se uma revisão sistemática, sob forma de artigo publicado, evidenciando a falta de estudos utilizando a AM em conjunto com a AGA para previsões de qualquer tipo. Em seguida realizamos uma revisão da literatura como forma de ter uma visão geral sobre o tema.

Também apresentou-se uma abordagem para tratamento de dados deste contexto com objetivo de maximizar a performance de algoritmos de AM. Detalhou-se todas as etapas e precauções tomadas durante o pré-processamento para que não ocorresse vazamento de dados, prejudicando o desempenho.

Em seguida, realizou-se dois ciclos de DSR onde, no primeiro ciclo, identificou-se quais variáveis mais importantes e as minimamente necessárias para prever o desfecho "óbito em 6 meses" (Quadro 28), sendo um achado importante para otimizar trabalhos futuros. Ainda no primeiro ciclo, avaliamos um conjunto de sete algoritmos de AM em suas configurações-padrão.

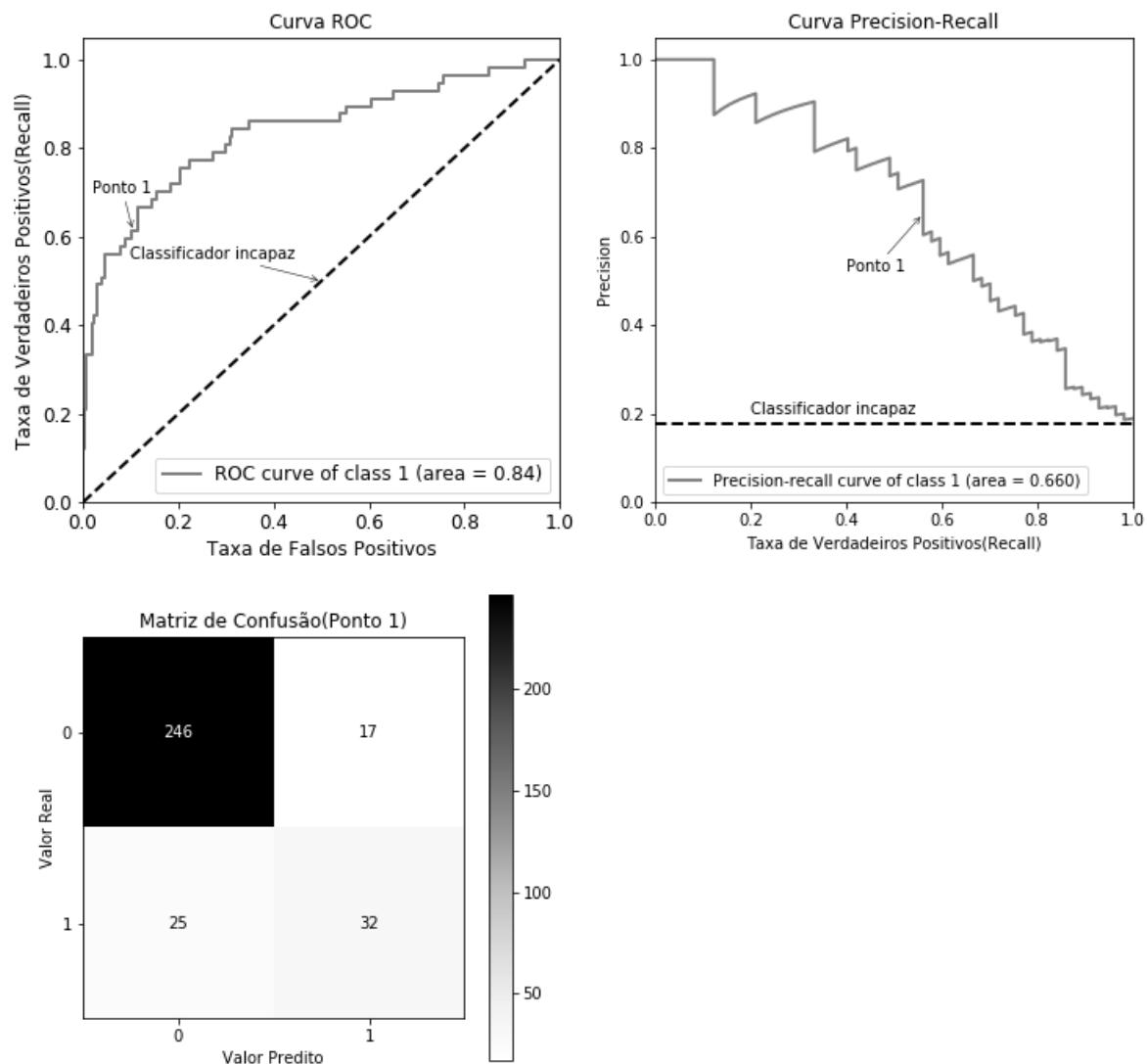
No segundo ciclo, realizamos uma otimização dos algoritmos em busca do melhor desempenho, considerando a métrica *AUC precision-recall*. O classificador de melhor desempenho utilizou o algoritmo Máquina de Vetores de Suporte (SVM), que obteve *AUC Precision-Recall* igual a 0,660 e conseguiu identificar 32 dos 57 pacientes de alto risco da base de testes (dados não vistos). As curvas ROC, *Precision-Recall* bem como a matriz de confusão estão na Figura 24.

Retornando a pergunta inicial da pesquisa:

**Como classificar os pacientes oncológicos em alto ou baixo risco de óbito em 6 meses utilizando as variáveis presentes na AGA?**

**Resposta:** A partir do modelo de AM supervisionada desenvolvido com o algoritmo máquina de vetores de suporte (SVM), utilizando os parâmetros **C = 1.7** e **kernel = rbf**, em conjunto com as 43 *features* selecionadas como minimamente necessárias e otimizado para detecção de pacientes da classe minoritária (alto risco de óbito). Este algoritmo obteve *AUC Precision-Recall* igual a 0,660 e *AUC-ROC* igual a 0,836.

Figura 24 – Curvas ROC, Precision-Recall e matriz de confusão (exemplo) para o modelo desenvolvido utilizando o algoritmo SVM.



Fonte: O autor

Analizando o objetivo geral e os específicos:

- **Objetivo geral:** Desenvolver um modelo de AM capaz de classificar os pacientes idosos em tratamento oncológico em baixo e alto risco de morte em 6 meses, avaliando os dados disponíveis nos exames que compõem a (AGA) e outras variáveis disponíveis.

Atendeu-se ao objetivo geral com sucesso visto que criou-se um modelo de AM com desempenho esperado. O algoritmo selecionado foi o máquina de vetores de suporte (SVM), conforme demonstrado no texto.

**Objetivos específicos:**

1. Determinar qual o subconjunto de variáveis preditoras (*features*) capaz de realizar a classificação com a mesma qualidade que o conjunto completo.

Determinamos um sub-conjunto de apenas 43 variáveis com capacidade de predição superior ao conjunto completo, de 344 variáveis. Resultado repetido no quadro 28:

É importante destacar que o modelo, mesmo com bom poder de classificação e bem calibrado, não diz o que deve ser feito pela equipe médica para modificar o desfecho. Analisando as variáveis, identificamos que elas se dividem em dois grupos distintos: as não modificáveis e as modificáveis, detalhadas abaixo:

- **Não modificáveis:** As do primeiro grupo são pertencentes as dimensões **tipo de câncer, hábitos pregressos e estado mental**. O tipo de Câncer ([61]-Câncer de próstata) ou o histórico de consumo de bebida alcoólica (Etil-Ja-Bebeu-Quanto-Tempo), por exemplo, embora sejam utilizadas no modelo, não podem ser alteradas.
- **Modificáveis:** As variáveis deste grupo pertencem as demais dimensões do quadro 28. São, por exemplo, a **avaliação nutricional, qualidade de vida, resultados clínicos e polifarmácia**. Essas variáveis podem ser alteradas pela equipe multidisciplinar de acompanhamento para reverter o desfecho.

Quadro 28 – Resumo do número de variáveis necessárias no algoritmo de AM por dimensão

Dimensão	Total de variáveis necessárias (RFE)
Nutrição	16
Qualidade de vida	13
Resultados clínicos	4
Mobilidade	4
Estado Geral	1
Estado Mental	2
Tipo de Câncer	1
Polifarmácia	1
Hábitos pregressos	1
Total	43

Fonte: O autor.

2. Avaliar, de um conjunto de algoritmos de aprendizagem de máquina, o que apresenta o melhor desempenho.

Avaliamos um conjunto de sete algoritmos de AM, conforme descrito no texto.

3. Otimizar a performance do algoritmo selecionado.

Várias ações listadas abaixo contribuíram para a melhora da performance dos algoritmos:

- Pré-processamento das variáveis de entrada.
- Seleção de variáveis.
- Otimização de hiperparâmetros.

#### 4. Priorizar a identificação dos pacientes de alto risco.

Utilizou-se a métrica **precisão** para otimizar os algoritmos de AM desenvolvidos e a métrica *AUC Precision-Recall* como critério de escolha. Ambas são métricas para a classe positiva (minoritária) que, no nosso contexto, priorizam a identificação dos pacientes de alto risco de óbito.

## 8.2 RESULTADOS ALCANÇADOS

Como resultados alcançados, listamos:

- O modelo de AM desenvolvido utilizando o algoritmo SVM mostrou-se o mais adequado dentre sete avaliados.
- Identificou-se o escopo mínimo de 43 variáveis necessárias para a elaboração do modelo de AM.
- Determinou-se uma sequencia de pré-processamento adequada aos dados deste contexto.

## 8.3 AMEAÇAS À VALIDADE

Os dados foram coletados em apenas um centro, o IMIP, portanto há esse importante viés a ser considerado. Seria importante dispor de uma base de dados maior e mais representativa, incluindo outros centros, de forma a desenvolver um algoritmo de AM com menor viés e, portanto, melhor desempenho.

## 8.4 LIMITAÇÕES DA PESQUISA

Utilizamos os algoritmos de AM mais utilizados e identificados na nossa revisão sistemática, não esgotando todas as classes de algoritmos existentes. Notadamente, não nos aprofundamos no design de redes neurais por entender ser uma área de pesquisa por si só. Outro limitante importante foi o tamanho da base de dados disponível para estudo (1600 pacientes).

## 8.5 TRABALHOS FUTUROS

Como trabalhos futuros, consideramos as seguintes oportunidades:

- Completar o ciclo de engenharia, implementando o modelo de AM desenvolvido. Uma proposta é integrá-lo ao aplicativo CONEXÃO VIDA, que foi desenvolvido na fábrica de software (chamada *Infinity*) que fiz parte enquanto estudante do mestrado e que se encontra em uso no IMIP.
- Utilizar os novos dados coletados no IMIP para realimentar o modelo de AM.
- Estudar novos modelos de AM utilizando diferentes arquiteturas de redes neurais e redes neurais profundas (*deep learning*).
- Avaliar o uso de outros algoritmos de balanceamento para bases de dados, notadamente o *borderline SMOTE* (HAN; WANG; MAO, 2005).

## REFERÊNCIAS

- AGRAWAL, A. et al. A lung cancer outcome calculator using ensemble data mining on SEER data. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2011. Citado 2 vezes nas páginas 47 e 49.
- ALI, A. et al. Analyzing potential of SVM based classifiers for intelligent and less invasive breast cancer prognosis. *2010 2nd International Conference on Computer Engineering and Applications, ICCEA 2010*, v. 2, p. 313–319, 2010. Citado 2 vezes nas páginas 46 e 49.
- ALJAWAD, D. A. et al. Breast cancer surgery survivability prediction using Bayesian network and support vector machines. *2017 International Conference on Informatics, Health and Technology, ICIHT 2017*, IEEE, v. 8, p. 1–6, 2017. Citado 2 vezes nas páginas 47 e 49.
- ALTMAN, N. S. *Practical Statistics for Medical Research*. EUA: Chapman & Hall/CRC., 1990. Citado na página 30.
- ALTMAN, N. S. An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, Taylor Francis, v. 46, n. 3, p. 175–185, 1992. Disponível em: <<https://www.tandfonline.com/doi/abs/10.1080/00031305.1992.10475879>>. Citado na página 37.
- ANDERSON, F. et al. Palliative performance scale (pps): a new tool. 1996. Citado na página 25.
- BARSAINYA, A.; SAIRAM, A.; PATIL, A. P. Analysis and Prediction of Survival after Colorectal Chemotherapy using Machine Learning Models. *2018 International Conference on Advances in Computing, Communications and Informatics, ICACCI 2018*, IEEE, p. 862–865, 2018. Citado 2 vezes nas páginas 48 e 49.
- BARTHOLOMAI, J. A.; FRIEBOES, H. B. Lung Cancer Survival Prediction via Machine Learning Regression, Classification, and Statistical Techniques. *2018 IEEE International Symposium on Signal Processing and Information Technology, ISSPIT 2018*, IEEE, p. 632–637, 2019. Citado 2 vezes nas páginas 48 e 49.
- BELLMAN, R. *Adaptive Control Processes*. EUA: Princeton Legacy Library, 1961. Citado na página 33.
- BEYER, A.; LANEY, D. The importance of ‘big data’: a definition,” gartner research report. 2012. Citado na página 19.
- BGS. The BGS toolkit for comprehensive geriatric assessment in primary care settings. 2018. Disponível em: <<https://www.bgs.org.uk/sites/default/files/content/resources/files/2018-08-23/CGAinPrimaryCareSettings.pdf>>. Citado 3 vezes nas páginas 21, 24 e 25.
- BISHOP, C. *Pattern Recognition and Machine Learning*. Springer, 2006. Disponível em: <<https://www.microsoft.com/en-us/research/publication/pattern-recognition-machine-learning/>>. Citado 4 vezes nas páginas 27, 28, 37 e 38.

- BISHOP, C. M. *Neural Networks for Pattern Recognition*. 1º. ed. USA: Oxford University Press, 1995. Citado na página 58.
- BRAY, F. et al. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, v. 68, n. 6, p. 394–424, 2018. Citado 2 vezes nas páginas 19 e 20.
- BROWNLEE, J. *Master Machine Learning Algorithms - Discover how they work and implement them from scratch*. [s.n.], 2016. Disponível em: <<https://machinelearningmastery.com/master-machine-learning-algorithms/>>. Citado 5 vezes nas páginas 35, 36, 37, 38 e 39.
- BROWNLEE, J. Machine Learning Mastery with Python. 2019. Citado 9 vezes nas páginas 30, 31, 35, 36, 37, 38, 39, 40 e 72.
- CHARLSON, M.; POMPEI, P.; ALES, K. M. C. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. 1987. Citado na página 25.
- CHAWLA, N. V. et al. Smote: synthetic minority over-sampling technique. 2002. Citado na página 35.
- CHEN, D. et al. A clustering approach in developing prognostic systems of cancer patients. *Proceedings - 7th International Conference on Machine Learning and Applications, ICMLA 2008*, n. ii, p. 723–728, 2008. Citado 2 vezes nas páginas 46 e 49.
- CHEN, G. M. et al. Prediction of survival in patients with liver cancer using artificial neural networks and classification and regression trees. *Taiwan Journal of Public Health*, v. 30, n. 5, p. 481–493, 2011. ISSN 10232141. Citado 2 vezes nas páginas 46 e 49.
- CHEN, T.; GUESTRIN, C. Xgboost: A scalable tree boosting system. 2016. Citado na página 39.
- CHIAVEGATTO, A. D. P. Uso de big data em saúde no Brasil : perspectivas para um futuro próximo \*. v. 24, n. 2, p. 325–332, 2015. Citado na página 19.
- CHIAVEGATTO, A. D. P. Machine Learning aplicado à Saúde. Workshop: Machine Learning. *19º Simpósio Brasileiro de Computação Aplicado à Saúde. Sociedade Brasileira de Computação*, 2019. Disponível em: <<https://sol.sbc.org.br/livros/index.php/sbc/catalog/view/29/95/245-1>>. Citado na página 31.
- CHIAVEGATTO, A. D. P. et al. Machine learning for predictive analyses in health: An example of an application to predict death in the elderly in São Paulo, Brazil. *Cadernos de Saude Publica*, v. 35, n. 7, p. 1–16, 2019. Citado 2 vezes nas páginas 29 e 57.
- CIRKOVIC, B. R. et al. Prediction models for estimation of survival rate and relapse for breast cancer patients. *2015 IEEE 15th International Conference on Bioinformatics and Bioengineering, BIBE 2015*, 2015. Citado 2 vezes nas páginas 47 e 49.
- COUNCIL, N. R. *Toward Precision Medicine: Building a Knowledge Network for Biomedical Research and a New Taxonomy of Disease*. Washington, DC: The National Academies Press, 2011. ISBN 978-0-309-22222-8. Disponível em: <<https://www.nap.edu/catalog/13284/>>

- toward-precision-medicine-building-a-knowledge-network-for-biomedical-research>. Citado na página 19.
- COX, M.; ELLSWORTH, D. Application-controlled demand paging for out-of-core visualization. 1997. Citado na página 19.
- CRESWELL, J. *Research design: qualitative, quantitative and mixed methods approaches*. [S.l.: s.n.], 2014. Citado na página 51.
- CUETO-LÓPEZ, N. et al. A comparative study on feature selection for a risk prediction model for colorectal cancer. *Computer Methods and Programs in Biomedicine*, v. 177, p. 219–229, 2019. ISSN 18727565. Citado na página 76.
- DRESCH LACERDA, J. e. a. *Design science research: método de pesquisa para avanço da ciência e tecnologia*. [S.l.: s.n.], 2015. Citado na página 54.
- ELFIKY, A. A. et al. Development and Application of a Machine Learning Approach to Assess Short-term Mortality Risk Among Patients With Cancer Starting Chemotherapy. *JAMA network open*, v. 1, n. 3, p. e180926, 2018. ISSN 2574-3805. Citado 2 vezes nas páginas 48 e 49.
- ENGSTRÖM, E.; STOREY, M.; RUNESON, P. How software engineering research aligns with design science: a review. *Empirical Software Engineering*, 2020. Citado na página 52.
- EXTERMANN, M. Geriatric Oncology: An Overview of Progresses and Challenges. *Cancer Research and Treatment*, 2010. ISSN 1598-2998. Citado na página 21.
- EXTERMANN, M.; HURRIA, A. Comprehensive geriatric assessment for older patients with cancer. *Journal of Clinical Oncology*, v. 25, n. 14, p. 1824–1831, 2007. Citado 2 vezes nas páginas 20 e 21.
- EYAL, N.; LAST, M.; RUBIN, E. Comparison of three classifiers for breast cancer outcome prediction. p. 1–6, 2015. Citado 2 vezes nas páginas 47 e 49.
- FAYERS, P. et al. The eortc qlq-c30 scoring manual. 2001. Citado na página 25.
- FERNÁNDEZ, A. et al. *Learning from Imbalanced Data Sets*. [S.l.: s.n.], 2018. Citado na página 42.
- FOLSTEIN, M. F.; FOLSTEIN, S. E.; MCHUGH, P. "mini-mental state". a practical method for grading the cognitive state of patients for the clinician. 1975. Citado na página 25.
- FU, D.; CHENG, Z.; DING, J. Applying machine learning in cancer prognosis using expression profiles of candidate genes. *ACM International Conference Proceeding Series*, Part F143213, p. 124–127, 2018. Citado 2 vezes nas páginas 48 e 49.
- GAN, B.; ZHENG, C. H.; WANG, H. Q. A survey of pattern classification-based methods for predicting survival time of lung cancer patients. *Proceedings - 2014 IEEE International Conference on Bioinformatics and Biomedicine, IEEE BIBM 2014*, n. 1408085, p. 5–12, 2014. Citado 2 vezes nas páginas 47 e 49.

- GANGGAYAH, M. D. et al. Predicting factors for survival of breast cancer patients using machine learning techniques. *BMC Medical Informatics and Decision Making*, BMC Medical Informatics and Decision Making, v. 19, n. 1, p. 1–17, 2019. ISSN 14726947. Citado 2 vezes nas páginas 48 e 49.
- GARM, A.; PARK, G. H.; SONG, X. Using an Electronic Comprehensive Geriatric Assessment and Health Coaching to Prevent Frailty in Primary Care: The CARES Model. *Medical Clinical Reviews*, n. October, 2018. Citado na página 21.
- GUIGOZ, Y. The mini nutritional assessment(mna) - review of the literature - what does it tell us? 2006. Citado 2 vezes nas páginas 25 e 78.
- HAN, H.; WANG, W.; MAO, B. Borderline-smote: a new over-sampling method in imbalanced data sets learning. 2005. Citado na página 99.
- HAN, J.; KAMBER, M. *Data Mining: Concepts and Techniques*. [S.l.: s.n.], 2000. ISBN 1-55860-489-8. Citado na página 34.
- HAN, J.; KAMBER, M. *Data Mining: Concepts and Techniques*. [S.l.: s.n.], 2001. Citado na página 34.
- HE, H.; GARCIA., E. Learning from imbalanced data. knowledge and data engineering. 2009. Citado na página 35.
- HEVNER, A. R. et al. *Design Science in Information Systems Research. MIS Quarterly*. [S.l.: s.n.], 2004. Citado 3 vezes nas páginas 54, 56 e 57.
- HINKLE, D.; WIERSMA, W.; JURS, S. *Applied Statistics for the Behavioral Sciences*. [S.l.: s.n.], 2003. Citado na página 30.
- HIRT, R.; NIKLAS, K.; SATZGER, G. An End-to-End Process Model for Supervised Machine Learning Classification : From Problem to Deployment in Information Systems. *Proceedings of the Conference on Design Science Research in Information Systems and Technology (DESRIST)*, n. May, 2017. Citado 7 vezes nas páginas 60, 61, 63, 72, 73, 74 e 86.
- HOWLADER, N. et al. *SEER Cancer Statistics Review, 1975-2017, National Cancer Institute. Bethesda, MD*. 2020. Disponível em: <[https://seer.cancer.gov/csr/1975\\_2017/](https://seer.cancer.gov/csr/1975_2017/)>. Citado na página 20.
- JAJROUDI, M. et al. Prediction of survival in thyroid cancer using data mining technique. *Technology in Cancer Research and Treatment*, v. 13, n. 4, p. 353–359, 2014. ISSN 1533-0346. Citado 2 vezes nas páginas 47 e 49.
- JAYASURYA, K. et al. Comparison of Bayesian network and support vector machine models for two-year survival prediction in lung cancer patients treated with radiotherapy. *Medical Physics*, v. 37, n. 4, p. 1400–1407, 2010. ISSN 00942405. Citado 2 vezes nas páginas 46 e 50.
- JOCHEMS, A. et al. A prediction model for early death in non-small cell lung cancer patients following curative-intent chemoradiotherapy. *Acta Oncologica*, Informa UK Limited, trading as Taylor & Francis Group, v. 57, n. 2, p. 226–230, 2018. ISSN 1651-226X. Citado 2 vezes nas páginas 48 e 49.

- KARADAGHY, O. A. et al. Development and Assessment of a Machine Learning Model to Help Predict Survival among Patients with Oral Squamous Cell Carcinoma. *JAMA Otolaryngology - Head and Neck Surgery*, v. 66160, p. 1–6, 2019. ISSN 21686181. Citado 2 vezes nas páginas 48 e 49.
- KARHADE, A. V. et al. Development of Machine Learning Algorithms for Prediction of 5-Year Spinal Chordoma Survival. *World Neurosurgery*, Elsevier Inc, v. 119, p. e842–e847, 2018. ISSN 1878-8769. Citado 2 vezes nas páginas 48 e 49.
- KARNOFSKY, d. B. J. The clinical evaluation of chemotherapeutic agents in cancer. 1949. Citado na página 25.
- KATZ, S. et al. Progress in development of the index of adl. 1970. Citado na página 25.
- KAUFMAN, S.; ROSSET, S.; PERLICH, C. Leakage in data mining: Formulation, detection, and avoidance. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, p. 556–563, 2011. Citado 2 vezes nas páginas 31 e 32.
- KENIS, C.; WILDERS, H. *Pratice Guidelines - Comprehensive Geriatric Assessement (CGA) in oncological patients*. [S.l.: s.n.], 2011. Citado 3 vezes nas páginas 21, 24 e 66.
- KOHAVI, R.; JOHN, H. Artificial Intelligence Wrappers for feature subset selection. v. 97, n. 97, p. 273–324, 1997. Citado na página 29.
- KOUROU, K. et al. Machine learning applications in cancer prognosis and prediction. *Computational and Structural Biotechnology Journal*, Elsevier B.V., v. 13, p. 8–17, 2015. ISSN 20010370. Disponível em: <<http://dx.doi.org/10.1016/j.csbj.2014.11.005>>. Citado na página 19.
- KUMARI, N. M. J.; KISHORE, K. K. V. Prognosis of diseases using machine learning algorithms: A survey. 2018. Citado 3 vezes nas páginas 40, 42 e 72.
- KURGAN, L. A.; MUSILEK, P. A survey of Knowledge Discovery and Data Mining process models. *Knowledge Engineering Review*, v. 21, n. 1, p. 1–24, 2006. Citado na página 60.
- KURSA, M. B.; RUDNICKI, W. R. Feature selection with the boruta package. *Journal of Statistical Software*, v. 36, n. 11, p. 1–13, 2010. ISSN 15487660. Citado na página 31.
- LACERDA, T. B. et al. Machine learning applied to survival prediction of elderly cancer patients : Systematic review. 2019. Citado na página 21.
- LEMAÎTRE, G.; NOGUEIRA, F.; ARIDAS, C. K. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of Machine Learning Research*, v. 18, p. 1–5, 2017. ISSN 15337928. Citado na página 35.
- LISBOA, P. J. G. et al. Time-to-event analysis with artificial neural networks: An integrated analytical and rule-based study for breast cancer. *IEEE International Conference on Neural Networks - Conference Proceedings*, v. 21, p. 2533–2538, 2008. ISSN 10987576. Citado 2 vezes nas páginas 46 e 50.

- LODHI, M. K. et al. A framework to predict outcome for cancer patients using data from a nursing EHR. *Proceedings - 2016 IEEE International Conference on Big Data, Big Data 2016*, p. 3387–3395, 2016. Citado 2 vezes nas páginas 47 e 49.
- LUO, W. et al. Guidelines for developing and reporting machine learning predictive models in biomedical research: A multidisciplinary view. *Journal of Medical Internet Research*, v. 18, n. 12, 2016. ISSN 14388871. Citado 2 vezes nas páginas 29 e 69.
- MATHIAS, S.; NAYAK, U. S.; ISAACS, B. Balance in elderly patients: the "get-up and go" test. 1986. Citado na página 25.
- MEDEIROS, A.; LACERDA, T.; BATISTA, R. Conexão vida: A proposal to assist in the treatment and communication between patients and imip medical staff. p. 1–4, June 2019. ISSN 2166-0727. Citado na página 19.
- MEI, X. Predicting five-year overall survival in patients with non-small cell lung cancer by reliefF algorithm and random forests. *Proceedings of 2017 IEEE 2nd Advanced Information Technology, Electronic and Automation Control Conference, IAEAC 2017*, p. 2527–2530, 2017. Citado 2 vezes nas páginas 47 e 49.
- MOHILE, S. G. et al. A pilot study of the vulnerable elders survey-13 compared with the comprehensive geriatric assessment for identifying disability in older patients with prostate cancer who receive androgen ablation. *Cancer*, v. 109, 2007. ISSN 0008543X. Citado na página 21.
- MUKAKA, M. M. Statistics corner : A guide to appropriate use of correlation coefficient in medical research. 2012. Citado 2 vezes nas páginas 30 e 66.
- NAM, Y.; SHIN, H. A Hybrid Cancer Prognosis System Based on Semi-Supervised Learning and Decision Trees. p. 640–648, 2013. Citado 2 vezes nas páginas 47 e 49.
- OLIVEIRA, M. M. de et al. Estimativa de pessoas com diagnóstico de câncer no Brasil: dados da Pesquisa Nacional de Saúde, 2013. *Revista Brasileira de Epidemiologia*, v. 18, n. suppl 2, p. 146–157, 2016. Citado na página 78.
- OVERCASH, J. et al. Comprehensive Geriatric Assessment as a Versatile Tool to Enhance the Care of the Older Person Diagnosed with Cancer. p. 1–13, 2019. Citado na página 24.
- OVERCASH, J. A. et al. The abbreviated comprehensive geriatric assessment (aCGA) for use in the older cancer patient as a prescreen: Scoring and interpretation. *Critical Reviews in Oncology/Hematology*, 2006. ISSN 10408428. Citado na página 21.
- OWUSU, C.; BERGER, N. A. Comprehensive geriatric assessment in the older cancer patient: Coming of age in clinical cancer care. *Clinical Practice*, v. 11, n. 6, p. 749–762, 2014. ISSN 20449046. Citado 2 vezes nas páginas 20 e 21.
- PEFFERS, K. et al. *A design science research methodology for information systems research*. [S.l.: s.n.], 2007. Citado na página 54.
- POURHOSEINGHOLI, M. A.; KHEIRIAN, S.; ZALI, M. R. Comparison of basic and ensemble data mining methods in predicting 5-year survival of colorectal cancer patients. *Acta Informatica Medica*, v. 25, n. 4, p. 254–258, 2017. ISSN 1986-5988. Citado 2 vezes nas páginas 47 e 49.

- PRATI, R. C.; BATISTA, G. E.; MONARD, M. C. Data mining with imbalanced class distributions: concepts and methods. 2009. Citado na página 34.
- PYLE, D. *Data Preparation for Data Mining*. [S.l.: s.n.], 1999. Citado na página 51.
- PYTHON. *Python*. 2020. Disponível em: <<http://www.python.org>>. Acesso em: 01 mai 2020. Citado na página 58.
- RAMJAUN, A.; NASSIF, M. O.; KROTNEVA, S. Improved targeting of cancer care for older patients: A systematic review of the utility of comprehensive geriatric assessment. *Journal of Geriatric Oncology*, Elsevier Inc., v. 4, n. 3, p. 271–281, 2013. Disponível em: <<http://dx.doi.org/10.1016/j.jgo.2013.04.002>>. Citado na página 21.
- RASTGOO, M. et al. Tackling the problem of data imbalancing for melanoma classification. 2016. Citado na página 35.
- RIPLEY, B. D. *Pattern Recognition and Neural Networks*. [S.l.: s.n.], 1996. Citado na página 58.
- ROSENBLATT, F. The perceptron: A probabilistic model for information storage and organization in the brain. 1958. Citado na página 38.
- RUDNICKI, W. R. et al. A statistical method for determining importance of variables in an information system. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, v. 4259 LNAI, p. 557–566, 2006. ISSN 16113349. Citado na página 31.
- SAFIYARI, A.; JAVIDAN, R. Predicting lung cancer survivability using ensemble learning methods. *2017 Intelligent Systems Conference, IntelliSys 2017*, v. 2018-Janua, n. September, p. 684–688, 2018. Citado 2 vezes nas páginas 47 e 49.
- SAILER, F. et al. Prediction of 5-year survival with data mining algorithms. *Studies in Health Technology and Informatics*, v. 213, p. 75–78, 2015. ISSN 18798365. Citado 2 vezes nas páginas 47 e 49.
- SAS. *Frequently Asked Questions about Neural Networks*. 2017. Eletronic. Disponível em: <<ftp://ftp.sas.com/pub/neural/FAQ.html>> Citado na página 58.
- SCIKIT-CEARN. 2020. Disponível em: <<http://scikit-learn.org>>. Acesso em: 01 mai 2020. Citado 6 vezes nas páginas 35, 36, 37, 38, 39 e 58.
- SCIPY. *SciPy*. 2020. Disponível em: <<http://www.scipy.org>>. Acesso em: 01 mai 2020. Citado na página 58.
- SENA, G. R. *Predição de óbito precoce em pacientes idosos com câncer por meio de aprendizagem de máquina*. 2016. Citado na página 50.
- SENA, G. R. et al. Developing Machine Learning Algorithms for the Prediction of Early Death in Elderly Cancer Patients : Usability Study Corresponding Author :. v. 5, p. 1–10, 2019. Citado 3 vezes nas páginas 48, 49 e 50.
- SEPEHRI, K. et al. A Computerized Frailty Assessment Tool at Points-of-Care: Development of a Standalone Electronic Comprehensive Geriatric Assessment/Frailty Index (eFI-CGA). *Frontiers in Public Health*, v. 8, n. March, p. 1–14, 2020. ISSN 2296-2565. Citado na página 21.

- SETHI, A. Analogizing of Evolutionary and Machine Learning Algorithms for Prognosis of Breast Cancer. *2018 7th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)*, IEEE, p. 252–255, 2018. Disponível em: <<https://ieeexplore.ieee.org/document/8748502/>>. Citado 2 vezes nas páginas 48 e 50.
- SHAHBAZIAN, N.; JAFARI, R. M.; HAGHNIA, S. Predictive model for survival in patients with gastric cancer. *Electronic Physician*, v. 8, n. 10, p. 3057–3061, 2016. Citado 2 vezes nas páginas 47 e 49.
- SHUKLA, N. et al. Breast cancer data analysis for survivability studies and prediction. *Computer Methods and Programs in Biomedicine*, v. 155, p. 199–208, 2018. ISSN 18727565. Citado na página 48.
- SHUKLA, N. et al. Breast cancer data analysis for survivability studies and prediction. *Computer Methods and Programs in Biomedicine*, Elsevier Ireland Ltd, v. 155, p. 199–208, 2018. ISSN 1872-7565. Citado na página 49.
- SMOLA, A. J.; SCHÖLKOPF, B. *A tutorial on support vector regression*. 2004. Citado na página 35.
- STRUNKIN, D.; NAMEE, B. M.; KELLEHER, J. D. An investigation into feature selection for oncological survival prediction. *Proceedings of the 9th International Conference on Information Technology, ITNG 2012*, p. 764–768, 2012. Citado 2 vezes nas páginas 47 e 49.
- VAISHNAVI, V. K.; KUECHLER, W. *Design Science Research Methods and Patterns: Innovating Information and Communication Technology*. 2<sup>a</sup>. ed. EUA: CRC Press, 2015. Citado na página 55.
- VERMA, M. Personalized medicine and cancer. 2012. Citado na página 19.
- WALCZAK, S.; CERPA, N. Heuristic principles for the design of artificial neural networks. *Information and Software Technology*, v. 41, n. 2, p. 107–117, 1999. ISSN 09505849. Citado 2 vezes nas páginas 39 e 80.
- WANG, Y. et al. A tree ensemble-based two-stage model for advanced-stage colorectal cancer survival prediction. *Information Sciences*, Elsevier Inc., v. 474, p. 106–124, 2019. ISSN 0020-0255. Citado 2 vezes nas páginas 48 e 49.
- WELSH, T. Comprehensive geriatric assessment—a guide for the non-specialist. 2014. Citado na página 21.
- WIERINGA, R. J. *Design science methodology: For information systems and software engineering*. 1º. ed. The Netherlands: Springer, 2014. 1-332 p. ISBN 9783662438398. Citado 5 vezes nas páginas 52, 53, 54, 55 e 56.
- WILDERS, H. et al. International society of geriatric oncology consensus on geriatric assessment in older patients with cancer. *Journal of Clinical Oncology*, v. 32, n. 24, p. 2595–2603, 2014. ISSN 15277755. Citado 3 vezes nas páginas 20, 21 e 24.
- WU, D. et al. An examination of TNM staging of melanoma by a machine learning algorithm. *ICCH 2012 Proceedings - International Conference on Computerized Healthcare*, p. 120–126, 2012. Citado 2 vezes nas páginas 47 e 50.

- XGBOOST. *XGBoost*. 2020. Disponível em: <<https://xgboost.readthedocs.io/en/latest/parameter.html>>. Acesso em: 01 mai 2020. Citado na página 40.
- YANG, Q.; WU., X. 10 challenging problems in data mining research. 2006. Citado 2 vezes nas páginas 34 e 35.
- YATES, D. S.; MOORE, D. S.; STARNES, D. S. *The Practice of Statistics*. 2<sup>a</sup>. ed. EUA: Freeman & Company, W&H, 2003. Citado na página 32.
- YAU, T. O. Precision treatment in colorectal cancer: Now and the future tung. 2019. Citado na página 19.
- YESAVAGE, J. A.; AL. et. Development and validation of a geriatric depression screening scale: a preliminary report. 1982. Citado na página 25.
- ZHANG, W.; TANG, J.; WANG, N. Using the machine learning approach to predict patient survival from high-dimensional survival data. *Proceedings - 2016 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2016*, p. 1234–1238, 2017. Citado 2 vezes nas páginas 47 e 49.

## Apêndices

APÊNDICE A – ARTIGO: MACHINE LEARNING  
APPLIED TO SURVIVAL PREDICTION OF  
ELDERLY CANCER PATIENTS: SYSTEMATIC  
REVIEW

# Machine Learning Applied to survival prediction of elderly cancer patients: Systematic Review

Tiago Beltrão Lacerda, Alberto Medeiros, Regis Batista Perez  
<sup>1</sup>CESAR School  
Recife Center for Advanced Studies and Systems  
Recife, Pernambuco, Brazil  
[tblacerda@outlook.com](mailto:tblacerda@outlook.com), {albertomedeiros214,  
regisbatistaperez}@gmail.com,

Ana Paula Cavalcanti Furtado<sup>1,2</sup>  
<sup>2</sup>Computing Department  
Universidade Federal Rural de Pernambuco  
Recife, Pernambuco, Brazil  
[anapaula.furtado@ufrpe.br](mailto:anapaula.furtado@ufrpe.br)

**Abstract** — Machine Learning (ML) is being successfully used in many science areas, medicine being no exception. On the other hand, cancer is a heterogeneous disease consisting of various subtypes and possible treatments that generate a lot of data (Big Data). This study sets out to identify, evaluate and interpret published research that examines how predicting the outcome of treating cancer can benefit from making use of ML. To achieve this, a systematic review of the literature was conducted. This review resulted in finding 1,855 studies, 32 of which were identified as primary studies. They were then classified according to research area and the aspect of ML they focus on. The results show gaps in current research as no studies were identified on ML using Comprehensive Geriatric Assessment (CGA), a fundamental tool that is used to improve caring for elderly patients with cancer.

**Keywords** - Machine Learning; Cancer; Prediction; Survival; CGA; Systematic review

## I. INTRODUCTION

According to GLOBOCAN [3], the Global Cancer Observatory, cancer is a worldwide, public health problem. GLOBOCAN estimates that by 2030, the annual figures will be: 27 million incident cases, 17 million deaths and 75 million people with cancer. The greatest number of people affected will be in low- and middle-income countries. Therefore, for all these reasons, it is a very relevant problem for study.

Cancer is a disease that tends to become more prevalent as people get older. It is expected to be the biggest cause of deaths worldwide in the coming years and this number is expected to increase as life expectancy increases [4]. The incidence of cancer is directly linked to chronological age, but age alone does not determine a person's state of health. To aid this understanding, the Comprehensive Geriatric Assessment (CGA) was developed as an instrument of interdisciplinary and multidimensional diagnostic assessment. It assesses the medical, psychological and functional capabilities of elderly patients in order to draw up an integrated treatment plan and long-term follow-up [5].

Machine learning (ML), an area of Artificial Intelligence (AI), deals with algorithms which are able to learn, find patterns and make predictions based on previous data [6]. It has been used extensively in the medical field, in conjunction with large

amounts of social, psychological, genetic, medical, treatment and other routines, which are collected and can be stored [7].

A search on the scientific portal IEEE Xplorer for the strings "machine learning" AND "cancer" covering the period from January 2008 to mid-August 2019 resulted in 1,583 articles being found, 478 of which have been published since 2018. This demonstrates that interest in this topic has been growing rapidly. bearing in mind that this portal is not dedicated to medicine.

This study is an effort to map the current state of knowledge about using ML to predict cancer survival, the focus being on predicting outcomes in elderly patients based on medical data. This paper is structured as follows: this Introduction presents the basic concepts related to the theme. In the next section, the methods, procedures and protocols used in the systematic review are presented while in section 3, the results and analysis of the data are set out in detail. Finally, in the last section, we draw some conclusions and make proposals for future lines of research.

## II. APPLYING A PROTOCOL

We adapted the Systematic Mapping for Software Engineering protocol described by Kitchenham and Charters [9] blended with Budgen [10], to define the protocol for the systematic review of the literature that was followed in this paper, according to the following steps:

- 1) Define topic and research questions.
- 2) Define inclusion and exclusion criteria.
- 3) Define search strings.
- 4) Select studies with peer review.
- 5) Read articles selected.
- 6) Conduct a Qualitative Analysis.

### A. Research question

The purpose of this review is to identify primary studies that use ML to classify elderly patients undergoing cancer treatment or to predict outcomes (survival time). To achieve this goal, we used the following question as the basis of the research:

- How can machine learning be applied to predict for how long elderly patients undergoing cancer treatment will survive?

To improve understanding, and to guide the review, secondary questions were drawn up to support the main subject research topic:

- What features can be used in the prediction?
- Is it possible to use information from the CGA to make predictions?
- How can elderly patients be categorized with respect to risk of death?
- What machine learning algorithms and techniques are used?
- What are the opportunities and challenges related to using machine learning and predicting survival in elderly patients undergoing cancer treatment?

#### *B. Establishing search strings in order to identifying articles*

In this study, the following search engines and databases with access within the academic environment were used:

- ACM (Association for Computing Machinery) Digital Library
- IEEE Xplore
- ScienceDirect – Elsevier
- PubMed
- Journals of Clinical Oncology

The search string below was used, and took account of synonyms:

- ("machine learning" OR "data mining" OR "data science") AND ("cancer") AND ("prediction")

The syntax may vary according to the search engine used; however, the broader search standard was used constantly, i.e., the one that found the most articles. The filter could be used in the title, metadata or even in the content of the article, when this option was available. The result obtained, after applying the search strings, is presented in Table I.

TABLE I. TOTAL STUDIES FOUND PER DIGITAL LIBRARY

Digital Library	Number of articles
ACM Digital Library	401
IEEE Xplore	734
ScienceDirect - Elsevier	226
PubMed	707
Journals of Clinical Oncology	54
<b>Total</b>	<b>2122</b>

The exclusion criteria used, when verifying if an article could be selected, were:

- Studies not identified by keywords (search terms);
- that were not in text format;
- that did not demonstrate empirical evidence;
- with partial data, that were still in progress or summaries;
- that were not scientific;
- that presented only personal opinions, even from experienced researchers in the field;
- that were not written in English;
- that were not accessible on the internet, nor on the search engines consulted;

#### *C. Applying the Protocol*

The first phase consisted of accessing the search engines available on the internet, within the academic environment, to apply the search strings in accordance with the appropriate syntax for each search engine. The result of this research was exported to *csv* files to enable greater mobility and data manipulation (including the removal of repeated articles). In the case of Sciencedirect the data were exported in *bib* format and later converted to *csv*. After doing so, a total of 2,122 articles was obtained, there being 1,856 unique articles after the repeated articles had been excluded.

In the second step, the titles of the articles were read in the light of the inclusion and exclusion criteria, and while always seeking answers to the research questions. Two researchers independently read each article, marking them '1' as approved or '0' as not approved for the next step. Soon afterwards, only the articles that the two researchers had marked '1' were selected. Thus, 228 articles were selected for analysis in the next step. Then, the same format was applied for reading the summary of the articles, with peer review, and this finally resulted in a total of 32 articles being selected. In Table II, we present the number of studies selected by each phase, as well as the rate of agreement between the two researchers.

TABLE II. NUMBER OF ARTICLES PER PHASE OF APPLICATION OF THE PROTOCOL.

Phase	Number of articles	Agreement Rate
Search on Digital Libraries	2122	-
Search on Digital Libraries (unique cases only)	1855	-
Analysis of the titles	0227	81.52%
Analysis of the abstracts	0032	77,.63%

#### *C. Quality analysis*

After the selection process described in previous sections, a careful reading of the studies selected was carried out and quality criteria were applied. Based on the models described by Kitchenam, Chatters [9] and Budgen [10], eight criteria compiled as questions were listed, this number being

considered adequate to classify the studies selected by the search filters.

- Q1: Does the study examine how ML can be used together with medical, social data and assessments, such as CGA, to predict results or classify cancer patients?
- Q2: Is the study based on research, and not just on expert opinion?
- Q3: Are the study's objectives clearly defined?
- Q4: Was the context of the study adequately described?
- Q5: Were the data collection methods correctly used and described?
- Q6: Was the research project adequately constructed to achieve the research objectives?
- Q7: Have the research results been properly validated?
- Q8: Does the study contribute to research or the daily needs of elderly citizens in any way?

The first criterion was proposed to assess whether the study was focused on exploring the use of machine learning applied to predicting the length of survival of cancer patients from the time that treatment started).

Criteria Q2 to Q7, on the other hand, measure the methods of the research: how close are they to studies that can be reproduced and follow written standards for validating scientific research results?

Question Q8 is intended to measure whether the study contributes to research or whether it is applicable to society in general or patients. In our case specifically, we assess whether any tools have been developed for fieldwork, be it a *webpage* or a *smartphone* application.

The results of applying the quality criteria are shown in Table III. This step, within the protocol adopted in the research, does not eliminate any of the articles, but may well indicate which of them are likely to be the most used in the discussion of the results.

TABLE III. QUALITATIVE ANALYSIS OF ARTICLES

Ref	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Total
[12]	1	1	1	1	1	1	1	1	8
[13]	1	1	1	1	1	1	1	1	8
[14]	1	1	1	1	1	1	1	0	7
[15]	1	1	1	1	1	1	1	0	7
[16]	1	1	1	1	1	1	1	0	7
[17]	1	1	1	1	1	1	1	0	7
[18]	1	1	1	1	0	0	1	1	6
[8]	1	1	1	1	1	1	1	0	6
[19]	1	1	1	1	0	1	1	0	6
Total	30	32	30	25	8	24	22	3	

[20]	1	1	1	1	0	1	1	0	6
[21]	1	1	1	1	0	1	1	0	6
[22]	1	1	1	1	0	1	1	0	6
[23]	1	1	1	1	0	1	1	0	6
[24]	1	1	1	1	0	1	1	0	6
[25]	1	1	1	1	0	1	1	0	6
[26]	1	1	1	1	0	1	1	0	6
[27]	1	1	1	1	0	1	1	0	6
[28]	0	1	1	1	0	1	1	0	5
[29]	1	1	1	0	0	1	1	0	5
[30]	1	1	1	0	0	1	1	0	5
[31]	1	1	1	0	0	1	1	0	5
[32]	1	1	1	0	0	1	1	0	5
[33]	1	1	1	1	0	1	0	0	5
[34]	0	1	1	1	0	1	0	0	4
[35]	1	1	1	0	0	1	0	0	4
[36]	1	1	1	1	0	0	0	0	4
[37]	1	1	1	1	0	0	0	0	4
[38]	1	1	1	1	0	0	0	0	4
[39]	1	1	1	1	0	0	0	0	4
[40]	1	1	0	0	1	0	0	0	3
[41]	1	1	0	1	0	0	0	0	3
[11]	1	1	1	0	0	0	0	0	3
Total	30	32	30	25	8	24	22	3	

### III. RESULTS OBTAINED

The protocol proposed by Kitchenam, Chatters [9] and Budgen [10] was used to present and analyze the results jointly by the main topics which were identified and verified when the articles were read. As to the research question, emphasis was given to how the theme of *machine learning* was presented, this being dividing into the groups below:

#### A. Purpose of the article

Most studies focus on breast cancer (10), followed by lung cancer (8), colorectal cancer (4) and cancer in general (3). Table IV shows the complete list.

Breast cancer is the most common type of cancer worldwide and is the most common type of cancer in women. This is followed by lung cancer [15] and colorectal cancer [25] (which

affect both genders). It is observed, therefore, that the most recurrent types are also the ones that are the most studied.

TABLE IV. CANCER TYPE X NUMBER OF ARTICLES.

Cancer Type	N. of articles
Breast	10
of the lungs	8
Colorectal	4
Cancer in general	3
Chordoma, Stomach, Liver, Melanoma, Oral squamous cell carcinoma, Kidneys, Thyroid	1(each)
Total	32

Twenty-nine studies applied ML to a specific type of cancer and only three studied survival without making a distinction in the type of cancer. This proportion reflects the difficulty in developing a generalized ML algorithm for a wide variety of different cancers [6]. On the other hand, there is a difficulty in making predictions for a specific type of cancer due to the paucity of data, mainly in rarer types of the disease.

Most studies (16) propose predicting survival for a period of up to 5 years, while 5 studies make predictions for up to 5 years and also for other periods, of up to 10 years. There is a consensus that the farther in the future, the more difficult it is to make a correct prediction.

#### B. Total number of patients in study

The study by Cirkovic [36] used only 58 patients. At the other end of the spectrum, Ali, Khan and others [8] used 162,500 patients. On examining the mean (22,038), median (6,414) and standard deviations (38,886.1), it appears that there is no consensus on the minimum number of patients to include. As a rule, the more data there is, the better it is to develop an ML model. In practice, the data available for the problem in question is used. It is observed that the largest volumes of data were in studies carried out using public bases, such as *Surveillance, Epidemiology, and End Results* (SEER). SEER is the official repository of the American government and covers about 26% of the population. It is the largest public database of cancer treatment [42], the dataset of which includes demographic, clinical and histopathological data.

#### C. Country and year of publication

Most studies (13) were carried out in the United States of America, but there are studies from thirteen other countries, including China (4), Iran (3) and India (2). Germany, Canada, South Korea, France, the Netherlands, Israel, Malaysia, Russia, Serbia, and Taiwan from each of which there is one article.

Finally, half of the articles (16) were published in the last four years (2017 onwards), the other half between 2008 and 2016, which proves to be a very relevant finding.

#### D. Model Features

There was a great variation in the number of features used per study. Studies by Jayasurya [16], Aljawad [19], Chen [11] and Wu [41] used only 4 features each. On the other hand, Elfiky [14], Eyal [21] and Gan [22] used, respectively, 5,000, 1,800 and 11,000 features. The high number is justified using genetic data, as did Eyal [21] and Gan [22], or medical notes processed by AI [14].

Most studies (14) used Demographic, Clinical and Histopathological data as features for the models studied. Four studies, Eyal [21], Gan [22], Zhang [28] and Fu [29] used genetic data. To use these types of features, they point out that it is important to make an adequate selection of variables since, in the case of genetic data, there are usually thousands of features.

A study by Strunkin [33] used data from several sources, including text processing of each patient's medical notes. However, it is noted that genetic data were not used in conjunction with other types of data in any study.

#### E. Algorithms tested

The median of algorithms tested was 2. However, a study by Chen [12] tested 9 ML algorithms. Most frequently, namely in 12 studies, only one ML algorithm was tested. The algorithm most used was *Artificial Neural Networks* (ANN), which was used 25 times in tests in studies related to 7 different types of cancers. Table V presents a summary.

TABLE V. QUANTITATIVE ANALYSIS

Ref.	Cancer	Best Algorithm	Total patients	Number of Features
[24]	Carcinoma	Random Forest	33065	44
[18]	Chordoma	BPM	265	5
[20]	Colorectal	DT	25000	13
[39]	Colorectal	DT	1568	44
[25]	Colorectal	Random Forest	395	17
[26]	Colorectal	Random Forest	14133	7
[35]	Liver	ANN	227	6
[38]	Gastric	Logistic Regression	227	15
[40]	General	DT	3685	10
[14]	General	Gradient Boosted Tree	23641	5390
[28]	General	NB	32	57
[27]	Breast	ANN	85189	29
[30]	Breast	ANN	2535	5
[36]	Breast	ANN	58	20
[15]	Breast	DT	8066	113
[32]	Breast	DT	5000	16
[34]	Breast	KNN	1466	34
[21]	Breast	Probabilistic Neural Network	1367	10
[19]	Breast	SVM	306	4
[29]	Breast	SVM	816	45
[8]	Breast	SVM	162500	16

[41]	Melanoma	EACCD	35952	4
[17]	Lungs	AdaBoost	10442	19
[16]	Lungs	BN	322	4
[12]	Lungs	J48	57254	63
[11]	Lungs	KNN	90214	4
[22]	Lungs	KNN	11868	11868
[13]	Lungs	Random Forest	10442	15
[23]	Lungs	Random Forest	1540	7
[31]	Lungs	Random Forest	5123	10
[33]	Kidneys	Voting feature interval	843	25
[37]	Thyroid	ANN	7706	16

#### IV. CONCLUSIONS

Of the 32 studies analyzed, the *random Forest* algorithm was the one with the best performance in 6 of them, three of them for evaluations related to lung cancer (Bartholomai [13], Jochems [23] and Mei [31]) followed by *Decision Trees* (DT) and *Artificial Neural Networks* (ANN) which had the best performance in 5 studies each, as can be seen in Table V.

The main objective of this study was to conduct research and an analysis to map the current state of knowledge about using ML to predict cancer survival, and using medical data to focus on predicting outcomes in elderly patients. To achieve this goal, we carried out a systematic review of the literature that covered more than 2,000 articles, which, by applying a protocol, as explained in the body of the article, led to identifying a total of 32 articles for discussion of topics such as: type of cancer under study, algorithms used, number of features and type of features used.

#### V. FUTURE WORK

In this review, no study was identified using CGA data with ML models of any type. We understand this to be an important gap to be studied, which we intend to fill by conducting a further study in near future.

#### REFERENCES

- [1] Vérende Dougoud-Chauvin, Jaé Jin Lee, Edgardo Santos, Vonetta L. Williams, Nicolo M.L. Battisti, Kavita Ghia, et al. "Using Big Data in oncology to prospectively impact clinical patient care: A proof of concept study". *Journal of Geriatric Oncology*, 9(6):665–672, 2018.
- [2] Janine Overcash, Nikki Ford, Elizabeth Kress, Caitlin Ubbing, and Nicole Williams. "Comprehensive Geriatric Assessment as a Versatile Tool to Enhance the Care of the Older Person Diagnosed with Cancer". pages 1–13, 2019.
- [3] Global cancer observatory. <https://gco.iarc.fr/>. Accessed: 2019-11-06.
- [4] Freddie Bray, Jacques Ferlay, Isabelle Soerjomataram, Rebecca L. Siegel, Lindsey A. Torre, and Ahmedin Jemal. "Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries". *CA: A Cancer Journal for Clinicians*, 68(6):394–424, 2018.
- [5] Hans Wildiers, Pieter Heeren, Martine Puts, Eva Topinkova, Maryska L.G. Janssen Heijnen, Martine Extermann, et al. "International Society of Geriatric Oncology consensus on geriatric assessment in older patients with cancer". *Journal of Clinical Oncology*, 32(24):2595–2603, 2014.
- [6] Konstantina Kourou, Themis P. Exarchos, Konstantinos P. Exarchos, Michalis V. Karamouzis, and Dimitrios I. Fotiadis. "Machine learning applications in cancer prognosis and prediction". *Computational and Structural Biotechnology Journal*, 13:8–17, 2015.
- [7] Alberto Medeiros. "Conexão vida: A proposal to assist in the treatment and communication between patients and IMIP medical staff". pages 1–4, June 2019.
- [8] \* Amna Ali, Umer Khan, Ali Tufail, and Minkoo Kim. "Analyzing potential of SVM based classifiers for intelligent and less invasive breast cancer prognosis". *2010 2nd International Conference on Computer Engineering and Applications*, ICCEA 2010, 2:313–319, 2010.
- [9] B. Kitchenham and S. Charters. issue: EBSE 2007-001. Technical report, 2(3), 2007.
- [10] David Budgen. "Performing Systematic Literature Reviews in Software Engineering", 2006.
- [11] \* Dechang Chen, Donald Henson, Arnold M. Schwartz, Kai Xing, Li Sheng, and Xiuzehen Cheng. "A clustering approach in developing prognostic systems of cancer patients". *Proceedings - 7th International Conference on Machine Learning and Applications*, ICMLA 2008, (ii):723–728, 2008.
- [12] \* Ankit Agrawal, Sanchit Misra, Ramanathan Narayanan, Lalith Polepeddi, and Alok Choudhary. "A lung cancer outcome calculator using ensemble data mining on SEER data". *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2011.
- [13] \* James A. Bartholomai and Hermann B. Friebes. "Lung Cancer Survival Prediction via Machine Learning Regression, Classification, and Statistical Techniques". *2018 IEEE International Symposium on Signal Processing and Information Technology*, ISSPIT 2018, pages 632–637, 2019.
- [14] \* Aymen A. Elfiky, Maximilian J. Pany, Ravi B. Parikh, and Ziad Obermeyer. "Development and Application of a Machine Learning Approach to Assess Short term Mortality Risk Among Patients With Cancer Starting Chemotherapy". *JAMA network open*, 1(3):e180926, 2018.
- [15] \* Mogana Darshini Ganggayah, Nur Aishah Taib, Yip Cheng Har, Pietro Lio, and Saninder Kaur Dhillon. "Predicting factors for survival of breast cancer patients using machine learning techniques". *BMC Medical Informatics and Decision Making*, 19(1):1–17, 2019.
- [16] \* K. Jayasurya, G. Fung, S. Yu, C. Dehing Oberije, D. De Ruysscher, A. Hope, W. De Neve, Y. Lievens, et al. "Comparison of Bayesian network and support vector machine models for two-year survival prediction in lung cancer patients treated with radiotherapy". *Medical Physics*, 37(4):1400–1407, 2010.
- [17] \* Ali Safiyari and Reza Javidan. "Predicting lung cancer survivability using ensemble learning methods". *2017 Intelligent Systems Conference, IntelliSys 2017*, 2018-January(September):684–688, 2018.
- [18] \* Aditya V. Karhade, Quirina Thio, Paul Oginik, Jason Kim, Santiago Lozano Calderon, Kevin Raskin, et al. "Development of Machine Learning Algorithms for Prediction of 5-Year Spinal Chordoma Survival". *World Neurosurgery*, 119:e842–e847, 2018.
- [19] \* Dania Abed Aljawad, Ebtesam Alqahtani, Ghaidaa Al Kuhaili, Nada Qamhan, Noof Alghamdi, Saleh Alrashed, et al. "Breast cancer surgery survivability prediction using Bayesian network and support vector machines". *2017 International Conference on Informatics, Health and Technology*, ICIHT2017, 8:1–6, 2017.
- [20] \* Aditya Barsainya, Anusha Sairam, and Annapurna P. Patil. "Analysis and Prediction of Survival after Colorectal Chemotherapy using Machine Learning Models". *2018 International Conference on Advances in Computing, Communications and Informatics*, ICACCI 2018, pages 862–865, 2018.
- [21] \* Noa Eyal, Mark Last, and Eitan Rubin. "Comparison of three classifiers for breast cancer outcome prediction". pages 1–6, 2015.
- [22] \* Bin Gan, Chun Hou Zheng, and Hong Qiang Wang. "A survey of pattern classification based methods for predicting survival time of lung cancer patients". *Proceedings 2014 IEEE International Conference on Bioinformatics and Biomedicine*, IEEE BIBM 2014, (1408085):5–12, 2014.
- [23] \* Arthur Jochems, Issam El-Naga, Marc Kessler, Charles S. Mayo, Shruti Jolly, Martha Matuszak, et al. "A prediction model for early death in non-small cell lung cancer patients following curative intent chemoradiotherapy". *Acta Oncologica*, 57(2):226–230, 2018.
- [24] \* Omar A. Karadaghy, Matthew Shew, Jacob New, and Andrés M. Bur. "Development and Assessment of a Machine Learning Model to Help Predict Survival among Patients with Oral Squamous Cell Carcinoma". *JAMA Otolaryngology - Head and Neck Surgery*, 66160:1–6, 2019.
- [25] \* Mohamad Amin Pourhoseingholi, Sedigheh Kheirian, and Mohammad Reza Zali. "Comparison of basic and ensemble data mining methods in predicting 5-year survival of colorectal cancer patients". *Acta Informatica Medica*, 25(4):254–258, 2017.
- [26] \* Fabian Sailer, Monika Pobiruchin, Sylvia Bochum, Uwe M. Martens, and Wendelin Schramm. "Prediction of 5-year survival with data mining algorithms". *Studies in Health Technology and Informatics*, 213:75–78, 2015.
- [27] \* Nagesh Shukla, Markus Hagenbuchner, Khin Than Win and Jack Yang. "Breast cancer data analysis for survivability studies and prediction". *Computer Methods and Programs in Biomedicine*, 155:199–208, 2018.
- [28] \* Wenbin Zhang, Jian Tang, and Nuo Wang. "Using the machine learning approach to predict patient survival from high-dimensional survival data". *Proceedings - 2016 IEEE International Conference on Bioinformatics and Biomedicine*, BIBM 2016, pages 1234–1238, 2017.
- [29] \* Danzhen Fu, Zijian Cheng, and Jiahao Ding. "Applying machine learning in cancer prognosis using expression profiles of candidate genes". *ACM International Conference Proceeding Series*, Part F143213:124–127, 2018.

- [30] \* Paulo J G Lisboa, Terence A. Etchells, Ian H. Jarman, M. S Hane Aung, Sylvie Chabaud, Thomas Bachelot, et al. "Time-to-event analysis with artificial neural networks: An integrated analytical and rule-based study for breast cancer". IEEE International Conference on Neural Networks - Conference Proceedings, 21:2533–2538, 2007.
- [31] \* Xueyan Mei. "Predicting five-year overall survival in patients with non-small cell lung cancer by reliefF algorithm and random forests". Proceedings of 2017 IEEE 2nd Advanced Information Technology, Electronic and Automation Control Conference, IAEAC 2017, pages 2527–2530, 2017.
- [32] \* Yonghyun Nam and Hyunjung Shin. "A Hybrid Cancer Prognosis System Based on Semi-Supervised Learning and Decision Trees". pages 640–648, 2013.
- [33] \* Dmitry Strunkin, Brian Mac Namee, and John D. Kelleher. "An investigation into feature selection for oncological survival prediction". Proceedings of the 9th International Conference on Information Technology, ITNG 2012, pages 764–768, 2012.
- [34] \* Anubha Sethi. "Analogizing of Evolutionary and Machine Learning Algorithms for Prognosis of Breast Cancer". 2018 7th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), pages 252–255, 2018.
- [35] \* G. Mei Chen, Chien Yeh Hsu, Hung Wen Chiu, B. A I Chyi-Huey, and W. U. Po-Hsun. "Prediction of survival in patients with liver cancer using artificial neural network and classification and regression trees". Taiwan Journal of Public Health, 30(5):481–493, 2011.
- [36] \* Bojana R. Andjelkovic Cirkovic, Aleksandar M. Cvetkovic, Srdjan M. Ninkovic, and Nenad D. Filipovic. "Prediction models for estimation of survival rate and relapse for breast cancer patients". 2015 IEEE 15th International Conference on Bioinformatics and Bioengineering, BIBE 2015, 2015.
- [37] \* M. Jajroudi, T. Baniasadi, L. Kamkar, F. Arbabi, M. Sanei, and M. Ahmadzade. "Prediction of survival in thyroid cancer using data mining technique". Technology in Cancer Research and Treatment, 13(4):353–359, 2014.
- [38] \* Nahid Shahbazian, Razieh Mohammad Jafari, and Sahar Haghnia. "Predictive model for survival in patients with gastric cancer. Electronic Physician, 8(10):3057–3061, 2016.
- [39] \* Yuyan Wang, Dujuan Wang, Xin Ye, Yanzhang Wang, Yunqiang Yin, and Yaochu Jin. "A tree ensemble based two-stage model for advanced stage colorectal cancer survival prediction". Information Sciences, 474:106–124, 2019.
- [40] \* Muhammad K. Lodhi, Rashid Ansari, Yingwei Yao, Gail M. Keenan, Diana J. Wilkie, and Ashfaq Khokhar. "A framework to predict outcome for cancer patients using data from a nursing EHR". Proceedings - 2016 IEEE International Conference on Big Data, Big Data 2016, pages 3387–3395, 2016.
- [41] \* Dengyuan Wu, Charles Yang, Stephen Wong, Jon Meyerle, Bowu Zhang, and Dechang Chen. "An examination of TNM staging of melanoma by a machine learning algorithm". ICCH 2012 Proceedings - International Conference on Computerized Healthcare, pages 120–126, 2012.
- [42] Seer. <https://seer.cancer.gov/about/>. Accessed: 2019-11-09.

## APÊNDICE B – LISTAGEM COMPLETA DAS VARIÁVEIS DA AGA REALIZADA NO IMIP

Atributo	Tipo	Range	Descrição
Anteced-Vacina-Toxoide-Tetanico	Categorica	[1] sim [2] não	Se recebeu vacina do Tétano
Anteced-Vacina-Pneumococica	Categorica	[1] sim [2] não	Se recebeu vacina Pneumocócica
Anteced-Vacina-Influenza	Categorica	[1] sim [2] não	Se recebeu vacina para Influenza
Anteced-Vacina-Hepat-B	Categorica	[1] sim [2] não	Se recebeu vacina para Hepatite-B Para analfabetos: corte em 19 pontos
MINI-MENTAL-Escore	Numérica	0..30	Com 1 a 3 anos de escolaridade: corte em 23 pontos Com 4 a 7 anos de escolaridade: corte em 24 pontos Acima de 7 anos de escolaridade: 28 pontos para se estabelecer o nível normal de cognição
MINI-MENTAL-Escore30	Categorica	0 a 30	Avaliação do estado cognitivo do paciente, aplica-se fatores de correção, de acordo com a escolaridade
Aval-GDS	Numérica	0..15	Escala de depressão geriátrica
Anteced-Hist-Quedas-Qtas	Categorica	maior ou igual a 0	Quantidade de quedas recentes
ATV-Fisica	Categorica	avaliar relevancia	avaliar relevancia
ATV-Fisica-Classif	Categorica	[1] Muito ativo, [2] Ativo, [3] Irregularmente Ativo, [4] Sedentário	Classificação de atividade física
Time-Up-Go	Numérica	acima de zero	Tempo para completar, em segundos
Time-Up-Go-Classi-Mobil	Categorica	[1] Normal, [2] Anormalidade Leve, [3] Anormalidade média, [4] Anormalidade moderada, [5] Anormalidade grave	Classificação quanto à mobilidade
Time-Up-Go-Classi-Ris-Qued	Categorica	[1] baixo risco de queda, [2] médio risco de queda, [3] alto risco de queda	Classificação quanto ao risco de queda
PPS	Numérica	0..100	Escala de performance Paliativa (PPS versão 2)
KATZ	Categorica	[0] Independente em todas as 6 funções, [1] dependente em 1 função, [2] dependente em 2 funções, [3] dependente em 3 funções, [4] dependente em 4 funções, [5] dependente em 5 funções, [6] dependente em todas 6 funções	Índice de KATZ – AVD
1-Infarto Mioc	Categorica	[1] Sim, [2] Não	Ocorrência de Infarto do miocárdio
1-Insuf Card Cong	Categorica	[1] Sim, [2] Não	Ocorrência de Insuficiência cardíaca congestiva
1-Doenca Vasc Perif	Categorica	[1] Sim, [2] Não	Ocorrência de doença vascular periférica
1-Doenca Cereb Vasc	Categorica	[1] Sim, [2] Não	Ocorrência de doença cerebrovascular
1-Doenca Pulm Cron	Categorica	[1] Sim, [2] Não	Ocorrência de doença pulmonar crônica
1-Doeenca Tecido Conj	Categorica	[1] Sim, [2] Não	Ocorrência de doença tecido conjuntivo
1-Diabete Leve S Complic	Categorica	[1] Sim, [2] Não	Ocorrência de diabetes leve, sem complicações
1-Ulcera Peptica	Categorica	[1] Sim, [2] Não	Ocorrência de úlcera péptica
2-Hemiplegia	Categorica	[1] Sim, [2] Não	Ocorrência de hemiplegia
2-Diabete C Complic	Categorica	[1] Sim, [2] Não	Ocorrência de diabete com complicação
2-Doenca Renal Sev Mod	Categorica	[1] Sim, [2] Não	Ocorrência de tumor
3-Doenca Fig Sev Mod	Categorica	[1] Sim, [2] Não	Ocorrência de leucemia
6-Tumor Malig Metas	Categorica	[1] Sim, [2] Não	Ocorrência de linfoma
6-SIDA	Categorica	[1] Sim, [2] Não	Ocorrência de doença do fígado severa ou moderada
0-HAS	Categorica	[1] Sim, [2] Não	Ocorrência de Hipertensão arterial sistêmica
CHARLSON	Numérica	0 .. 37	Índice de Comorbidade de Charlson
Cancer-Dsc	Categorica	-	Descrição do tipo de câncer conforme CID 10C
CID10	Categorica	-	código do CID10C
Mestastase	Categorica	[1] Ocorreu Metástase [2] Não ocorreu metástase	Ocorrência ou não de metástase
Metastase-Sítios	Categorica	[1] Óssea [2] Pulmonar [3] Sistema N. Central [4] Hepática [5] Outras	Código com local de ocorrência da metástase
Metastase-Sítios-Dsc	Categorica	-	Descrição do local de ocorrência da metástase
QLQ30-01	Categorica	[1] Não, [2] Pouco, [3] Moderadamente, [4] Muito	QLQ-C30 – Q1: Você tem dificuldade quando faz grandes esforços?
QLQ30-02	Categorica	[1] Não, [2] Pouco, [3] Moderadamente, [4] Muito	QLQ-C30 – Q2: Você tem dificuldade quando faz grandes caminhadas?
QLQ30-03	Categorica	[1] Não, [2] Pouco, [3] Moderadamente, [4] Muito	QLQ-C30 – Q3: Você tem qualquer dificuldade quando faz uma curta caminhada fora de casa?
QLQ30-04	Categorica	[1] Não, [2] Pouco, [3] Moderadamente, [4] Muito	QLQ-C30 – Q4: Você tem que ficar numa cama ou na cadeira durante o dia ?
QLQ30-05	Categorica	[1] Não, [2] Pouco, [3] Moderadamente, [4] Muito	QLQ-C30 – Q5: Você precisa de ajuda para se alimentar, se vestir, se lavar ou ir ao banheiro?
QLQ30-06	Categorica	[1] Não, [2] Pouco, [3] Moderadamente, [4] Muito	QLQ-C30 – Q6: Durante a última semana, tem sido difícil fazer suas atividades diárias?
QLQ30-07	Categorica	[1] Não, [2] Pouco, [3] Moderadamente, [4] Muito	QLQ-C30 – Q7: Durante a última semana, tem sido difícil ter atividades de divertimento ou lazer?
QLQ30-08	Categorica	[1] Não, [2] Pouco, [3] Moderadamente, [4] Muito	QLQ-C30 – Q8: Durante a última semana, teve falta de ar?

QLQ30-09	Categorica	[1] Não, [2] Pouco, [3] Moderadamente, [4] Muito	QLQ-C30 – Q9: Durante a última semana, você tem tido dor?
QLQ30-10	Categorica	[1] Não, [2] Pouco, [3] Moderadamente, [4] Muito	QLQ-C30 – Q10: Durante a última semana, você precisou repousar?
QLQ30-11	Categorica	[1] Não, [2] Pouco, [3] Moderadamente, [4] Muito	QLQ-C30 – Q11: Durante a última semana, você tem tido problemas para dormir?
QLQ30-12	Categorica	[1] Não, [2] Pouco, [3] Moderadamente, [4] Muito	QLQ-C30 – Q12: Durante a última semana, você tem se sentido fraco?
QLQ30-13	Categorica	[1] Não, [2] Pouco, [3] Moderadamente, [4] Muito	QLQ-C30 – Q13: Durante a última semana, você tem tido falta de apetite?
QLQ30-14	Categorica	[1] Não, [2] Pouco, [3] Moderadamente, [4] Muito	QLQ-C30 – Q14: Durante a última semana, você tem se sentido enjoado(a)?
QLQ30-15	Categorica	[1] Não, [2] Pouco, [3] Moderadamente, [4] Muito	QLQ-C30 – Q15: Durante a última semana, você têm vomitado?
QLQ30-16	Categorica	[1] Não, [2] Pouco, [3] Moderadamente, [4] Muito	QLQ-C30 – Q16: Durante a última semana, você tem tido prisão de ventre?
QLQ30-17	Categorica	[1] Não, [2] Pouco, [3] Moderadamente, [4] Muito	QLQ-C30 – Q: Durante a última semana, você tem tido diarréia?
QLQ30-18	Categorica	[1] Não, [2] Pouco, [3] Moderadamente, [4] Muito	QLQ-C30 – Q18: Durante a última semana, você esteve cansado(a)?
QLQ30-19	Categorica	[1] Não, [2] Pouco, [3] Moderadamente, [4] Muito	QLQ-C30 – Q19: Durante a última semana, a dor interfeiou em sua atividade diária?
QLQ30-20	Categorica	[1] Não, [2] Pouco, [3] Moderadamente, [4] Muito	QLQ-C30 – Q20: Durante a última semana, você tem tido dificuldade para se concentrar em coisas, com ler jornais ou ver televisão?
QLQ30-21	Categorica	[1] Não, [2] Pouco, [3] Moderadamente, [4] Muito	QLQ-C30 – Q21: Durante a última semana, você se sentiu nervoso(a)?
QLQ30-22	Categorica	[1] Não, [2] Pouco, [3] Moderadamente, [4] Muito	QLQ-C30 – Q22: Durante a última semana, você esteve preocupado(a)?
QLQ30-23	Categorica	[1] Não, [2] Pouco, [3] Moderadamente, [4] Muito	QLQ-C30 – Q23: Durante a última semana, você se sentiu irritado(a) facilmente?
QLQ30-24	Categorica	[1] Não, [2] Pouco, [3] Moderadamente, [4] Muito	QLQ-C30 – Q24: Durante a última semana, você se sentiu deprimido(a)?
QLQ30-25	Categorica	[1] Não, [2] Pouco, [3] Moderadamente, [4] Muito	QLQ-C30 – Q25: Durante a última semana, você tem tido dificuldade de se lembrar das coisas?
QLQ30-26	Categorica	[1] Não, [2] Pouco, [3] Moderadamente, [4] Muito	QLQ-C30 – Q26: Durante a última semana, a sua condição física ou o tratamento médico tem interferido em sua vida social?
QLQ30-27	Categorica	[1] Não, [2] Pouco, [3] Moderadamente, [4] Muito	QLQ-C30 – Q26: Durante a última semana, a sua condição física ou o tratamento médico tem interferido em suas atividades sociais?
QLQ30-28	Categorica	[1] Não, [2] Pouco, [3] Moderadamente, [4] Muito	QLQ-C30 – Q28QLQ-C30 – Q26: Durante a última semana, a sua condição física ou o tratamento médico tem lhe trazido dificuldades financeiras?
QLQ30-29	Categorica	[1] Péssima .. [7] Ótimo	QLQ-C30 – Q29: Como você classificaria a sua saúde geral, durante a última semana?
QLQ30-30	Categorica	[1] Péssima .. [7] Ótimo	QLQ-C30 – Q30: Como você classificaria a sua qualidade de vida geral, durante a última semana?
KARNOFSKY	Categorica	0 .. 100	Escala de Performance de Karnofsky
Hemoglobina-Resultado	Numérica	racional	Contagem Hemoglobina
Contag-Leucocitos-Resultado	Numérica	racional	Contagem de Laucócitos
Granulocitos-Resultado	Numérica	racional	Contagem de Granulócitos
Plaquetas-Resultado	Numérica	racional	Contagem de Plaquetas
Creatinina-Resultado	Numérica	racional	Nível de creatinina
Tabag-Q.01	Categorica	[0] não preenchido [1] sim, diariamente (ir para A e C) [2] sim, mas não diariamente(ir para B). [3] não	O paciente fuma atualmente?
Tabag-Q.01.a	Numérica	igual ou maior que zero	Quantos cigarros fuma por dia?
Tabag-Q.01.b	Numérica	igual ou maior que zero	Quantos cigarros fuma por semana?
Tabag-Q.01.c	Numérica	igual ou maior que zero [99] não lembra	Que idade tinha quando começou a fumar regularmente?
Tabag-Q.02	Categorica	[1] sim, diariamente [2] sim, mas não diariamente [3] não	No passado, o paciente fumou ?
Tabag-Q.02.a	Numérica	igual ou maior que zero	Qual a idade tinha quando começou a fumar regularmente(em anos)
Tabag-Q.02.b	Numérica	igual ou maior que zero	Qual idade tinha quando parou de fumar(em anos)?
Etil-Consume-Beb-Alcool	Numérica	[1] sim [2] não	Consome bebida alcoólica?
Etil-Frequencia	Categorica	Texto com frequencia de ingestão de alcool	dado não estruturado. Avaliar se mantem ou não
Etil-Quanto-Tempo	Categorica	Texto com histórico temporal	dado não estruturado. Avaliar se mantem ou não
Etil-Tipo-Bebida	Categorica	Texto com o tipo da bebida frequentemente ingerida	dado não estruturado. Avaliar se mantem ou não
Etil-Ja-Bebeu	Categorica	[1] sim [2] não	No passado, já bebeu regularmente?
Etil-Ja-Bebeu-Quanto-Tempo	Numérica	inteiro	Parou de beber a quanto tempo(em anos)
Etil-Ja-Bebeu-Dura-Qto-Tempo	Numérica	Maior ou igual a 0	Quantidade de anos em que o paciente alega ter bebido
Etil-Ja-Bebeu-Erm-Que-Otde	Categorica	Texto com alegada frequência de ingestão de alcool	dado não estruturado. Avaliar se mantem ou não
Etil-Ja-Bebeu-Tipo-Bebida	Categorica	Texto com o tipo da bebida frequentemente ingerida	dado não estruturado. Avaliar se mantem ou não
MAN-TR-A	Categorica	[0] Diminuição severa, [1] Diminuição moderada, [2] Sem diminuição	Nós ultimos três meses houve diminuição da ingesta alimentar devido à perda de apetite, problemas digestivos ou dificuldade para mastigar ou engolir?

MAN-TR-B	Categorica	[0] superior a 3 Kg, [1] não sabe informar, [2] entre 1 e 3 kg, [3] sem perda de peso	Perda de peso nos últimos meses
MAN-TR-C	Categorica	[0] Restrito ao leito ou à cadeira de rodas, [1] deambula mas não é capaz de sair de casa, [2] normal	Avaliação de Mobilidade(normal, cadeira de rodas ou leito)
MAN-TR-D	Categorica	[0] Sim, [2] Não	Passou por algum estresse psicológico ou doença aguda nos últimos três meses?
MAN-TR-E	Categorica	[0] Demência ou depressão leve, [1] Demência leve, [2] sem problemas psicológicos	Problemas neuropsicológicos
MAN-TR-F	Numérica	[0] IMC < 19, [1] 19<= IMC < 21, [2] 21<= IMC < 23, [3] IMC>= 23	IMC - Índice de massa corporea. Calculado dividindo-se o peso(kg) pela altura(m) ao quadrado.
MAN-Escore-TR	Numérica	0 .. 14	Sub escore da triagem. 12 pontos ou mais: normal. 11 pontos ou menos: possibilidade de desnutrição(continuar com a avaliação nutricional)
MAN-AvG-G	Categorica	[0] Não, [1] Sim	O Paciente vive em sua própria casa(não em casa geriátrica os hospital)
MAN-AvG-H	Categorica	[0] Não, [1] Sim	Utiliza mais de três medicamentos diferentes por dia
MAN-AvG-I	Categorica	[0] Não, [1] Sim	Lesões de pele ou escaras ?
MAN-AvG-J	Categorica	[0] uma refeição, [1] duas refeições, [2] três refeições	Quantas refeições faz por dia
MAN-AvG-K-1	Categorica	[0] Não, [1] Sim	Consume: pelo menos uma porção diária de leite ou derivados
MAN-AvG-K-2	Categorica	[0] Não, [1] Sim	Consume duas ou mais porções semanais de ovos ou legumes?
MAN-AvG-K-3	Categorica	[0] Não, [1] Sim	Consume carne, peixe ou aves todos os dias
MAN-AvG-K-4-Res	Categorica	[0] nenhuma ou uma resposta sim, [0.5] duas respostas sim, [1] três respostas sim	resumo das três perguntas anteriores
MAN-AvG-L	Categorica	[0] Não, [1] Sim	Consume duas ou mais porções diárias de frutas e vegetais?
MAN-AvG-M	Categorica	[0] menos de três, [0.5] três a cinco, [1] mais de cinco	Quantos copos de líquidos consome por dia?
MAN-AvG-N	Categorica	[0] não é capaz de se alimentar sozinho, [1] alimenta-se sozinho, porém com dificuldade, [2] alimenta-se sozinho sem dificuldade	Modo de se alimentar
MAN-AvG-O	Categorica	[0] acredita estar desnutrido, [1] não sabe dizer, [2] acredita não ter problema nutricional	O Paciente acreditar ter algum problema nutricional?
MAN-AvG-P	Categorica	[0] não muito boa, [0.5] não sabe informar, [1] boa, [2] melhor	Em comparação a outras pessoas da mesma idade, como o paciente considera a própria saúde?
MAN-AvG-Q	Numérica	[0] CB <21, [0.5] 21<= CB <22, [1] CB > 22	Circunferência do braço(CB) em cm
MAN-AvG-R	Numérica	[0] CP < 31, [1] CP >=31	Circunferência da panturrilha em cm
MAN-Aval-Global	Numérica	0 .. 16	Máximo de 16 pontos
MAN-Escore-Tot	Numérica	0 .. 30	Triagem + avaliação global
MAN-Aval-Est-Nut-Res	Categorica	[1] risco de desnutrição [2] desnutrido [3] sem problemas de nutrição	Avaliação final da nutrição
MAN-Aval-Est-Nutric	Numérica	0..99	Escore de nutrição
5-Polifarmacia-Medicamento	Numérica	maior ou igual a 0	número de medicamentos de uso continuo, receitados ou não.
REGISTRO	Registro	inteiro	Registro do Paciente na base do IMP. Chave primária entre as tabelas.
Sexo	Categorica	[M] [F]	Gênero do paciente
Idade(anos)	Numérica	>=60	Auto explicativo
Diagnostico Principal	Categorica	-	Descrição do diagnóstico principal, conforme CID10-C
Data Admissao	Numérica	-	Data de elaboração da AGA
Data Limite	Numérica	-	Data de Admissão + 180 dias.
Bairro	Categorica	-	Auto explicativo
Municipio	Categorica	-	Auto explicativo
Estado	Categorica	-	Auto explicativo
Q.01	Categorica		Escolaridade – Qual foi o último ano de estudo
Q.02	Categorica	[1] Solteiro, [2] Casado(a) legalmente, [3] União estável há mais de 6 meses, [4] viúvo(a), [5] Separado(a) Divorciado(a), [88] Não quis informar	Situação conjugal atual
Q.03	Categorica	[1] Aposentado com outra ocupação, [2] aposentado sem outra ocupação, [3] Trabalhos domésticos, [4] Trabalho fora do domicílio	Ocupação
Q.04	Categorica	[1] Católica, [2] Evangélica, [3] Espírita, [4] Budista, [5] Outra	descrição
Q.05	Numérica	-	Renda doméstica
Peso	Numérica	0..200	Peso em quilogramas
Altura	Numérica	1..3	Altura em metros
IMC	Numérica	0..30	Índice de massa Corporea = peso / (altura) <sup>2</sup>
Anteced-Hist-Quedas	Categorica	[1] sim [2] não	Se há histórico de quedas
Sobrevida(dias)	Numérica	>0	Dias corridos entre a data de Admissão e a Data de Obito, caso tenha ocorrido
Motivo da Saída	Saída	NA, ÓBITO	NA para os pacientes vivos

Data saída	Numérica	-	Data do óbito
DATANASC	Categorica	data	Data de nascimento
Hip-Diag-Cancer	Categorica	-	Descrição do tipo de câncer conforme CID 10C
Hip-Diag-CID10	Categorica	-	código do CID10C
06-Cor Pele	Categorica	[1] Branca [2] Negra [3] Amarela [4] Parda [5] Indígena [77] não sabe [88] não quis informar	Cor da pele do paciente

## Anexos

ANEXO A – ANEXO: FORMULÁRIO IMIP COM  
AGA E QUESTIONÁRIO SOCIO-ECONÔMICO

**FATORES DE RISCO PARA INFECÇÕES RELACIONADAS À ASSISTÊNCIA À SAÚDE (IrAS) EM PACIENTES ONCOLÓGICOS IDOSOS. UM ESTUDO DE COORTE PROSPECTIVA**

Nº CONTROLE PESQUISA: \_\_\_\_\_

*Checklist :*

Local etiqueta ou identificação do/a paciente

Nº	Escalas	Data	Responsável
1	Critérios de elegibilidade		
2	TCLE		
3	Variáveis sóciodemográficas		
4	Exames solicitados		
5	Polifarmácia	Avaliação Geriátrica Amplia	
6	Índice de Charlson		
7	Escala de Performance de Karnofsky		
8	MAN		
9	Mini mental		
10	GDS escala de depressão		
11	IPAQ		
12	Time up and go		
13	PPS		
14	Índice de Katz – AVD		
15	Qualidade de vida EORTC QL 30		
16	Definição da vulnerabilidade /AGA		
17	Resultados de exames / estadiamento		
18	Seguimento 1 – 30 dias		
19	Seguimento 2 – 60 dias		
20	Seguimento 3 – 90 dias		
21	Seguimento 4 – 120 dias		
22	Seguimento 5 – 150 dias		
23	Seguimento 6 – 180 dias		
24	Ficha de internamento /SPA		
25	Ficha de infecção – IrAS		
26			

IrAS: [1] Sim; [2] Não Data da 1ª IrAS: / /

Decisão para limitação da terapêutica (DLT): [1] Sim [2] Não Data da DLT: / /

Data da última da consulta: / /

Óbito : [1] Sim; [2] Não Data do óbito: / /

**Fluxograma do estudo:**

Atividade	*Recrutamento	Mês 1	Mês 2	Mês 3	Mês 4	Mês 5	Mês 6	Se internamento e/ou óbito Preencher ficha de seguimento específica
Aplicação de critérios de elegibilidade								
TCLE								
Ficha de admissão								
Avaliação geriátrica ampla (AGA)								
Qualidade de vida EORTC QL 30								
Hemograma								
Plaquetas								
Creatinina								
Perfil imunológico Toll Like								
Ficha de segmento mensal								
Avaliação de prontuário e sistema de informação hospitalar								

\* Preferencialmente no D1 -(admissão até início da primeira terapia – máximo:30 dias )

\*\* Avaliação geriátrica ampla (AGA): polifarmácia, Índice de Charlson, ECOG, Minimental, GDS, IPAQ, "time up and go", PPS, índice de Katz, Definição da Vulnerabilidade /AGA

## 1- CRITÉRIOS DE ELEGIBILIDADE

### CONTROLE DE PESQUISA:

Para todos os pacientes abordados para a pesquisa

### CRITÉRIOS DE INCLUSÃO

- Idade igual ou superior a 60 anos: [1]Sim [2] Não
- Capaz de aceitar autonomamente a participação no estudo: [1]Sim [2] Não
- Diagnóstico de câncer confirmado por:
  - Histologia: [1]Sim [2] Não
  - Citologia: [1]Sim [2] Não
  - Imunohistoquímica: [1]Sim [2] Não

### CRITÉRIOS DE EXCLUSÃO:

- Câncer de pele, tipo basocelular ou epidermóide não metastásico.  
[1]Sim [2] Não
- Pacientes com diagnóstico prévio de câncer, exceto câncer de pele, tipo basocelular ou epidermóide não metastásico  
[1]Sim [2] Não
- Pacientes com tratamento prévio de câncer, exceto cirurgia  
[1]Sim [2] Não

### Paciente participará da pesquisa?

- Não, não preenche critérios de elegibilidade
- Não, houve recusa em participar  
*se paciente não participa da pesquisa encaminhar para rotina*
- Sim  
**responder às questões seguintes**

## **2 – TERMO DE CONSENTIMENTO LIVRE E ESCLARECIDO (TCLE)**

Nome do projeto de pesquisa: **Fatores de risco para Infecções relacionadas à Assistência à Saúde (IrAS) em pacientes oncológicos idosos. Um estudo de coorte prospectiva.** (Tese de doutorado DINTER INCA/ IMIP)

**Pesquisadora:** Jurema Telles de Oliveira Lima

Pesquisadora principal/médica /oncologista clínica

CRM PE 11279

**Orientador:** Luiz Claudio Thuler

**Co-orientadora:** Maria Júlia Gonçalves de Mello

**Contato da pesquisadora:** Rua dos Coelhos 300 Boa Vista CEP 50070-550 telefone: 99763591 /21225731 email: jurema@imip.org.br

**Contato do Comitê de Ética em Pesquisa do IMIP:** Comitê de Ética e Pesquisa do IMIP- Rua dos Coelhos 300 Boa Vista CEP 50070-550. Prédio Administrativo Orlando Onofre do IMIP- 1º andar. Tel.: (81) 2122-4756 email: comitedeetica@imip.org.br

### **Voluntário**

Eu, Jurema Telles de Oliveira Lima, responsável pela pesquisa, faço parte de uma equipe e juntamente com este grupo, estamos fazendo um convite para o (a) senhor (a) participar como voluntário deste nosso estudo, porque o (a) senhor (a) é um(a) paciente com idade superior a sessenta anos que será acompanhado (a) no Serviço de Oncogeriatría do IMIP. Esta pesquisa pretende avaliar os fatores de risco para a ocorrência de infecção durante o seu tratamento. Acreditamos que a pesquisa seja importante porque pacientes com idade acima de sessenta anos, com o seu diagnóstico, realizando exames e/ou tratamentos podem ter um risco maior de desenvolver esta complicação (infecção), porém este risco varia muito com a situação do paciente e os diferentes tratamentos. A população brasileira está ficando cada vez mais idosa e há muitos poucos estudos sobre os melhores tipos de cuidados para esta população. Além do mais, cada pessoa desta população tem muitas características diferentes e essas diferenças podem ser importantes para o cuidado da saúde de cada um. Para a realização deste estudo, que já teve a autorização do comitê de ética do IMIP, será feito o seguinte: iremos acompanhar o seu atendimento no IMIP a partir da presente data até seis meses do seu tratamento. Não iremos modificar as condutas da equipe de saúde e de seu médico, apenas iremos observar seu acompanhamento, o resultado de exames e seus dados de prontuários. No início, será realizada uma entrevista, com duração de cerca de 20 minutos, para lhe conhecer melhor e /ou esclarecer

alguns dados de seu prontuário. Quando você for realizar um exame de sangue pedido por seu médico iremos aproveitar para coletar uma pequena amostra de sangue que permita fazer um exame de uma proteína que pode estar relacionada a este risco de desenvolver a infecção (avaliação da mutação dos receptores da família *toll like*).

Sua participação constará de permitir o acompanhamento de seu tratamento durante o período do estudo e atendimentos no IMIP. Conversaremos pessoalmente com você na sua admissão no estudo e se ocorrer uma infecção, e sempre que você precisar de informações e esclarecimentos, ou se ficarmos com alguma dúvida em seu prontuário. Os benefícios que esperamos desse estudo são conhecer a diversidade destes fatores de risco de modo a prevenir ainda mais a ocorrência de infecção e suas complicações, sem que atrapalhe o tratamento planejado e a qualidade de vida do paciente. Estes conhecimentos irão beneficiar os pacientes que irão enfrentar o tratamento e acompanhamento que você está passando agora, além de permitir que o IMIP e outros serviços planejem as ações de cuidados, principalmente para a pessoa mais idosa. É importante esclarecer que, caso você decida não participar, nada mudará em seu atendimento, sendo garantido o tratamento e acompanhamento de rotina previsto para você. Durante todo o período da pesquisa você tem o direito de tirar qualquer dúvida ou pedir qualquer outro esclarecimento, bastando para isso entrar em contato, com algum dos pesquisadores ou com o Conselho de Ética em Pesquisa do IMIP. Em caso de algum problema relacionado com a pesquisa você terá direito à assistência gratuita que será prestada no IMIP, como sempre. Você tem garantido o seu direito de não aceitar participar ou de retirar sua permissão, a qualquer momento, sem nenhum tipo de prejuízo ou retaliação, pela sua decisão. As informações desta pesquisa serão sempre confidenciais, e serão divulgadas apenas em eventos ou publicações científicas, não havendo identificação dos voluntários, a não ser entre os responsáveis pelo estudo, sendo assegurado o sigilo sobre sua participação. O sangue coletado será utilizado apenas para a finalidade científica desta pesquisa. Os gastos necessários para a sua participação na pesquisa serão assumidos pelos pesquisadores.

Exames de laboratório vão ser realizados com o seu sangue e com o material da biópsia. Com relação ao material da biópsia, um pequeno fragmento fresco será enviado para cultura celular e avaliação de infiltrado inflamatório e habitualmente tais testes são dispensáveis pois não alteram a conduta do seu tratamento, mas conhecê-los pode ser de utilidade para o futuro. Serão também realizadas coletas de 10 mL de sangue periférico para realização dos testes

laboratoriais que avaliarão a sua imunidade e também possíveis marcadores genéticos relacionados com sua doença, porém que não irão alterar o tipo do seu tratamento. O material excedente à rotina será armazenado no Laboratório de Pesquisa Translacional do IMIP, sob condições adequadas de temperatura para garantir a integridade das amostras e uso no projeto de pesquisa acima e o material de biópsia será guardado no Departamento de Anatomia Patológica do IMIP.

Existe um risco mínimo para o participante de pesquisa. A coleta poderá provocar uma mancha vermelha ou roxa no local da picada da agulha. Você e seus acompanhantes serão orientados como tratar, caso ocorra mancha vermelha ou roxa no local.

### CONSENTIMENTO

Acredito ter sido suficientemente informado a respeito das informações que li ou que foram lidas para mim, descrevendo o objetivo e os testes laboratoriais que serão realizados neste trabalho.

Ficaram claros para mim quais são os objetivos do estudo, os testes laboratoriais que serão realizados, seus desconfortos e riscos, as garantias de confidencialidade e de esclarecimentos permanentes. Ficou claro também estou isenta de despesas e que terei a garantia do acesso a tratamento hospitalar quando necessário. Concordei voluntariamente em participar deste estudo e poderei retirar o meu consentimento a qualquer momento, antes ou durante o mesmo, sem penalidades ou prejuízos ou perda de qualquer benefício que eu possa ter adquirido no meu atendimento neste serviço.

Estou ciente que minha participação é isenta de despesas ou ganhos financeiros e que isto não irá interferir no meu tratamento.

Se os meus materiais biológicos guardados no laboratório de pesquisa do IMIP forem utilizados em pesquisas aprovadas pelo Comitê de Ética em Pesquisa, escolho livremente a opção abaixo assinalada:

**Autorizo a utilização dos meus materiais biológicos sem necessidade de novo consentimento a cada pesquisa.**

[1] Sim [2] Não

**Desejo ser contatada para autorizar o uso dos meus materiais biológicos a cada pesquisa e em caso de impossibilidade de contato comigo, indico que a nova autorização seja fornecida pelo(a)a senhor (a)**

[1] Sim [2] Não

---

(nome e contato do familiar ou representante legal)

---

Nome e assinatura do paciente

Data \_\_\_\_/\_\_\_\_/\_\_\_\_

---

Assinatura do Responsável Legal/ testemunha imparcial Data \_\_\_\_/\_\_\_\_/\_\_\_\_

**Declaro que obtive de forma apropriada e voluntária o Consentimento Livre e Esclarecido deste paciente (ou de seu representante legal) para a cessão de material biológico para armazenamento no laboratório de Pesquisa Translacional acima referido.**

---

Nome e assinatura do responsável pela obtenção do termo

Data \_\_\_\_/\_\_\_\_/\_\_\_\_

---

Assinatura de uma testemunha

Data \_\_\_\_/\_\_\_\_/\_\_\_\_

---

Dados do pesquisado responsável:

Jurema Telles de Oliveira Lima

Pesquisadora principal/médica /Oncologista clínica - CRM 11279

Rua dos Coelhos 300 Boa Vista CEP 50070-550 Telefone: 99763591 /21224185 email: [jurema@imip.org.br](mailto:jurema@imip.org.br)

Comitê de ética e pesquisa do IMIP-Prédio Administrativo Orlando Onofre do IMIP- 1ºandar

Telefone 21224756

[comitedeetica@imip.org.br](mailto:comitedeetica@imip.org.br) IMIP

Local etiqueta ou identificação do/a paciente

**FICHA DE AVALIAÇÃO INICIAL**

Nº Controle da pesquisa: \_\_\_\_\_

Data do preenchimento \_\_\_\_/\_\_\_\_/\_\_\_\_\_

**3 – VARIÁVEIS SOCIODEMOGRÁFICAS E CLÍNICAS**

**IDENTIFICAÇÃO**

Nome: \_\_\_\_\_

Registro: \_\_\_\_\_ Data de Nascimento: \_\_\_\_ / \_\_\_\_ Idade: ( ) anos

Sexo: ( ) Masculino ( ) Feminino

Nome da mãe: \_\_\_\_\_

Endereço: \_\_\_\_\_ Complemento: \_\_\_\_\_

Bairro: \_\_\_\_\_ Município: \_\_\_\_\_ Estado: \_\_\_\_\_

Telefone para contato: \_\_\_\_\_

Cuidador: \_\_\_\_\_ Contato: \_\_\_\_\_

Qual seu maior desejo?

Caso você esteja numa situação que não possa tomar nenhuma decisão, quem você nomearia?

Contato: \_\_\_\_\_

**Escolaridade**

Qual foi o último ano de estudo concluído? \_\_\_\_\_  
(Ex : 6º ano do primeiro grau)

**Situação Conjugual atual:**

- [1] solteiro(a)
- [2] casado(a) legalmente
- [3] têm união estável há mais de seis meses
- [4] viúvo(a)
- [5] separado(a) ou divorciado(a)
- [88] não quis informar

**Ocupação**

- Aposentado com outra ocupação [1]
- Aposentado sem outra ocupação [2]
- Trabalhos domésticos [3]
- Trabalho fora do domicílio [4]
- Qual: \_\_\_\_\_
- Profissão quando trabalha : \_\_\_\_\_

**Religião**

- Católica [1]
- Evangélica [2]
- Espírita [3]
- Budista [4]
- Outra [5]

**Somando a renda de todas as pessoas que moram na sua casa, inclusive você, qual é o valor em reais:** \_\_\_\_\_

**A sua cor da pele é :**

- [1] branca
- [2] negra
- [3] amarela
- [4] parda
- [5] indígena
- [77] não sabe
- [88] não quis informar

**Você reside em área:**

- [1] Urbana
- [2] rural

**Quantas pessoas vivem dessa renda?** \_\_\_\_\_

Peso: \_\_\_\_\_ Altura: \_\_\_\_\_ IMC: \_\_\_\_\_

### QUEIXA PRINCIPAL

---



---



---

### HISTÓRIA DA DOENÇA ATUAL

---



---



---



---

#### ANTECEDENTES:

História de quedas no último ano: [1] Sim [2] Não Quantas? \_\_\_\_\_  
Onde? \_\_\_\_\_ Quando? \_\_\_\_\_ () Não sabe informar

Internamento recente (nos últimos 30 dias): [1] Sim [2] Não

Motivo: [ ] Clínico [ ] Cirúrgico Tempo: \_\_\_\_\_

Local: \_\_\_\_\_

Resumo de alta [1] Sim se Sim, anexar [2] Não

Cirurgia para retirada do baço: [1] Sim [2] Não

#### ANTECEDENTES VACINAIS:

Vacinas:

Toxóide tetânico ou dupla adulto				Outras vacinas		
1 <sup>a</sup> Dose	2 <sup>a</sup> Dose	3 <sup>a</sup> Dose	4 <sup>a</sup> Dose	Pneumocócica	Influenza Gripe	Hepatite B

### INTERROGATÓRIO SINTOMATOLÓGICO:

**Febre na atual doença** [1] Sim [2] Não Quando (início) \_\_\_\_/\_\_\_\_/\_\_\_\_

Quanto tempo \_\_\_\_\_

Aferiu a temperatura [1] Sim [2] Não Temperatura máxima \_\_\_\_\_ °C

## HÁBITOS PREGRESSOS

### Tabagismo

1. Atualmente, o (a) sr.(a) fuma?

- ( ) Sim, diariamente (ir para a e c)  
( ) Sim, mas não diariamente (ir para b)  
( ) Não

a. Quantos cigarros o(a) sr.(a) fuma por dia? \_\_\_\_\_

b. Quantos cigarros o(a) sr.(a) fuma por semana? \_\_\_\_\_ (apenas se Q = 2)

c. Que idade o (a) sr.(a) tinha quando começou a fumar regularmente? \_\_\_ anos  
[99] não lembra

2. No passado, o(a) sr.(a) já fumou? (se Sim responder sub- itens a e b)

- ( ) Sim, diariamente  
( ) Sim, mas não diariamente  
( ) Não

a. Que idade o(a) sr.(a) tinha quando começou a fumar regularmente? \_\_\_ anos  
[99] Não lembra

b. Que idade o(a) sr.(a) tinha quando parou de fumar? \_\_\_ anos [99] não lembra

### Etilismo:

Você consome bebida alcoólica? [1] Sim [2] Não

Com que frequência? \_\_\_\_\_ Quanto tempo? \_\_\_\_\_

Tipo de bebida? \_\_\_\_\_

Já bebeu? [1] Sim [2] Não

Se sim, parou há quanto tempo? \_\_\_ anos

Bebeu durante quanto tempo \_\_\_ anos; Em que quantidade? \_\_\_\_\_

Tipo de bebida? \_\_\_\_\_

Local etiqueta ou identificação do/a paciente

## HIPÓTESE DIAGNÓSTICA:

Câncer de \_\_\_\_\_ CID 10 : C\_\_\_\_\_

Metástases (conhecidas na admissão): [1] Sim [2] Não Data: \_\_\_ / \_\_\_ / \_\_\_  
 Sítios de metástases: [1]óssea [2]pulmonar [3] SNC [4]hepática [5]outras: \_\_\_\_\_

DATA	Sítio	Exame comprobatório do câncer *	Resultado

\* Exame comprobatório do câncer: [1] histologia [2] citologia [3] imunohistoquímica

## 4 – EXAMES SOLICITADOS

Exames relevantes para a pesquisa:

Exame	Data da solicitação
Hemograma com plaquetas	
Creatinina	
Perfil imunológico	

	Já realizados	Data de solicitação
Exames de estadiamento		

Local etiqueta ou identificação do/a paciente

**5 – POLIFARMÁCIA**

**MEDICAÇÃO E INFORMAÇÕES DE USO:**

Medicamentos de uso regular	Posologia	Efeitos Adversos (quais?)	Tempo de uso

Usa algum dos medicamentos citados acima por conta própria? [1] Sim [2] Não  
 Quais? \_\_\_\_\_

**POLIFARMÁCIA ( $\geq 5$  medicamentos)** [1] Sim [2] Não

Local etiqueta ou identificação do/a paciente

**6 – COMORBIDADES - ÍNDICE DE CHARLSON**

Peso	Condição Clínica	[1] Sim	[2] Não
1	<b>Infarto do miocárdio</b>	[1] Sim	[2] Não
	<b>Insuficiência cardíaca congestiva</b>	[1] Sim	[2] Não
	<b>Doença Vascular Periférica</b>	[1] Sim	[2] Não
	<b>Doença Cerebrovascular</b>	[1] Sim	[2] Não
	<b>Doença pulmonar crônica</b>	[1] Sim	[2] Não
	<b>Doença tecido conjuntivo</b>	[1] Sim	[2] Não
2	<b>Diabetes leve, sem complicações</b>	[1] Sim	[2] Não
	<b>Úlcera péptica</b>	[1] Sim	[2] Não
	<b>Hemiplegia</b>	[1] Sim	[2] Não
	<b>Diabete com complicações</b>		
	<b>Doença renal severa ou moderada</b>	[1] Sim	[2] Não
3	<b>Tumor</b>	[1] Sim	[2] Não
	<b>Leucemia</b>	[1] Sim	[2] Não
	<b>Linfoma</b>	[1] Sim	[2] Não
	<b>Doença do fígado severa ou moderada</b>	[1] Sim	[2] Não
6	<b>Tumor maligno, metástase</b>	[1] Sim	[2] Não
	<b>SIDA</b>	[1] Sim	[2] Não
0	<b>HAS</b>	[1] Sim	[2] Não
	<b>Outros</b> _____	[1] Sim	[2] Não

Escore: \_\_\_\_\_ (*Observação: não pontuar o tumor primário*)

Data: \_\_\_\_ / \_\_\_\_ / \_\_\_\_

Avaliador

**7– ESCALA DE PERFORMANCE KARNOFSKY**

<b>100%</b>	<b>Sem sinais ou queixas, sem evidência de doença</b>
<b>90%</b>	<b>Mínimos sinais e sintomas, capaz de realizar suas atividades com esforço</b>
<b>80%</b>	<b>Sinais e sintomas maiores, realiza suas atividades com esforço</b>
<b>70%</b>	<b>Cuida de si mesmo, não é capaz de trabalhar</b>
<b>60%</b>	<b>Necessita de assistência ocasional, capaz de trabalhar</b>
<b>50%</b>	<b>Necessita de assistência considerável e cuidados médicos frequentes</b>
<b>40%</b>	<b>Necessita de cuidados médicos especiais</b>
<b>30%</b>	<b>Extremamente incapacitado, necessita de hospitalização mas sem iminência de morte</b>
<b>20%</b>	<b>Muito doente, necessita suporte</b>
<b>10%</b>	<b>Moribundo, morte iminente</b>

Data: \_\_\_\_/\_\_\_\_/\_\_\_\_\_

Avaliador

Local etiqueta ou identificação do/a paciente

## 8 – AVALIAÇÃO NUTRICIONAL

### QUESTIONÁRIO MAN (miniavaliação nutricional)

Preencher a primeira parte deste questionário, indicando a resposta. Somar os pontos da Triagem. Caso o escore seja igual ou inferior a 11, concluir o questionário para obter a avaliação do estado nutricional.

#### Triagem

- A Nos últimos três meses houve diminuição da ingestão alimentar devido a perda de apetite, problemas digestivos ou dificuldade para mastigar ou deglutar?  
 0 = diminuição severa da ingestão  
 1 = diminuição moderada da ingestão  
 2 = sem diminuição da ingestão
- B Perda de peso nos últimos meses  
 0 = superior a três quilos  
 1 = não sabe informar  
 2 = entre um e três quilos  
 3 = sem perda de peso
- C Mobilidade  
 0 = restrito ao leito ou à cadeira de rodas  
 1 = deambula mas não é capaz de sair de casa  
 2 = normal
- D Passou por algum estresse psicológico ou doença aguda nos últimos três meses?  
 0 = sim      2 = não
- E Problemas neuropsicológicos  
 0 = demência ou depressão graves  
 1 = demência leve  
 2 = sem problemas psicológicos
- F Índice de massa corpórea (IMC = peso [kg] / estatura [m]<sup>2</sup>)  
 0 = IMC < 19  
 1 = 19 ≤ IMC < 21  
 2 = 21 ≤ IMC < 23  
 3 = IMC ≥ 23

**Escore de triagem** (subtotal, máximo de 14 pontos)    
 12 pontos ou mais      normal;  
                           desnecessário continuar a avaliação  
 11 pontos ou menos      possibilidade de desnutrição;  
                           continuar a avaliação

#### Avaliação global

- G O paciente vive em sua própria casa (não em casa geriátrica ou hospital)  
 0 = não      1 = sim
- H Utiliza mais de três medicamentos diferentes por dia?  
 0 = sim      1 = não
- I Lesões de pele ou escaras?  
 0 = sim      1 = não

Ref.: Guigoz Y, Vellas B and Garry PJ. 1994. Mini Nutritional Assessment: A practical assessment tool for grading the nutritional state of elderly patients. *Facts and Research in Gerontology*. Supplement # 2:15-59.  
 Rubenstein LZ, Harker J, Guigoz Y and Vellas B. Comprehensive Geriatric Assessment (CGA) and the MNA: An Overview of CGA, Nutritional Assessment, and Development of a Shortened Version of the MNA. In: "Mini Nutritional Assessment (MNA): Research and Practice in the Elderly". Vellas B, Garry PJ and Guigoz Y, editors. Nestlé Nutrition Workshop Series. Clinical & Performance Programme, vol. 1. Karger, Bâle, in press.

©1998 Société des Produits Nestlé S.A., Vevey, Switzerland, Trademark Owners

J Quantas refeições faz por dia?

- 0 = uma refeição  
 1 = duas refeições  
 2 = três refeições

K O paciente consome:

- pelo menos uma porção diária de leite ou derivados (queijo, iogurte)? sim  não
  - duas ou mais porções semanais de legumes ou ovos? sim  não
  - carne, peixe ou aves todos os dias? sim  não
- 0,0 = nenhum ou uma resposta «sim»  
 0,5 = duas respostas «sim»  
 1,0 = três respostas «sim»

 , 

L O paciente consome duas ou mais porções diárias de frutas ou vegetais?

- 0 = não      1 = sim

M Quantos copos de líquidos (água, suco, café, chá, leite) o paciente consome por dia?

- 0,0 = menos de três copos  
 0,5 = três a cinco copos  
 1,0 = mais de cinco copos

 , 

N Modo de se alimentar

- 0 = não é capaz de se alimentar sozinho  
 1 = alimenta-se sozinho, porém com dificuldade  
 2 = alimenta-se sozinho sem dificuldade

O O paciente acredita ter algum problema nutricional?

- 0 = acredita estar desnutrido  
 1 = não sabe dizer  
 2 = acredita não ter problema nutricional

P Em comparação a outras pessoas da mesma idade, como o paciente considera a sua própria saúde?

- 0,0 = não muito boa  
 0,5 = não sabe informar  
 1,0 = boa  
 2,0 = melhor

 , 

Q Circunferência do braço (CB) em cm

- 0,0 = CB < 21  
 0,5 = 21 ≤ CB ≤ 22  
 1,0 = CB > 22

 , 

R Circunferência da panturrilha (CP) em cm

- 0 = CP < 31      1 = CP ≥ 31

**Avaliação global** (máximo 16 pontos)

**Escore da triagem**

**Escore total** (máximo 30 pontos)

**Avaliação do Estado Nutricional**

de 17 a 23,5 pontos      risco de desnutrição

menos de 17 pontos      desnutrido

Data: \_\_\_\_\_ / \_\_\_\_\_ / \_\_\_\_\_

Avaliador

Local etiqueta ou identificação do/a paciente

## 9 – MINI MENTAL

Anos de estudo: [1] analfabeto [2] 1- 3 anos [3] 4 – 7anos [4] acima de 7 anos

Para os analfabetos considera-se o corte de 19 pontos; com 1 a 3 anos de escolaridade 23 pontos; 4 a 7 anos 24 pontos e acima de 7 anos de instrução 28 pontos para estabelecer nível normal de cognição.

### MINI-EXAME DO ESTADO MENTAL

(Folstein, Folstein & McHugh, 1975)

Paciente: \_\_\_\_\_

Data da Avaliação: \_\_\_\_ / \_\_\_\_ / \_\_\_\_ Avaliador: \_\_\_\_\_

#### ORIENTAÇÃO

- Dia da semana (1 ponto) .....( )
- Dia do mês (1 ponto) .....( )
- Mês (1 ponto) .....( )
- Ano (1 ponto) .....( )
- Hora aproximada (1 ponto) .....( )
- Local específico (aposento ou setor) (1 ponto) .....( )
- Instituição (residência, hospital, clínica) (1 ponto) .....( )
- Bairro ou rua próxima (1 ponto) .....( )
- Cidade (1 ponto) .....( )
- Estado (1 ponto) .....( )

#### MEMÓRIA IMEDIATA

- Fale 3 palavras não relacionadas. Posteriormente pergunte ao paciente pelas 3 palavras. Dê 1 ponto para cada resposta correta .....( )
- Depois repita as palavras e certifique-se de que o paciente as aprendeu, pois mais adiante você irá perguntá-las novamente.

#### ATENÇÃO E CÁLCULO

- (100 - 7) sucessivos, 5 vezes sucessivamente (1 ponto para cada cálculo correto) .....( )  
(alternativamente, soletrar MUNDO de trás para frente)

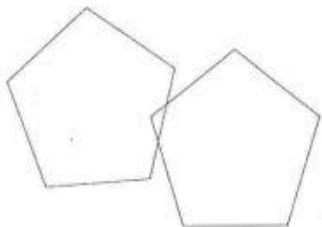
#### EVOCAÇÃO

- Pergunte pelas 3 palavras ditas anteriormente (1 ponto por palavra) .....( )

#### LINGUAGEM

- Nomear um relógio e uma caneta (2 pontos) .....( )
- Repetir "nem aqui, nem ali, nem lá" (1 ponto) .....( )
- Comando: "pegue este papel com a mão direita sobre ao meio e coloque no chão ( 3 pts) .....( )
- Ler e obedecer: "feche os olhos" (1 ponto) .....( )
- Escrever uma frase (1 ponto) .....( )
- Copiar um desenho (1 ponto) .....( )

ESCORE: ( \_\_\_ /30)



Data: \_\_\_\_ / \_\_\_\_ / \_\_\_\_

Avaliador

Local etiqueta ou identificação do/a paciente

**10 – GDS – ESCALA DE DEPRESSÃO  
GERIÁTRICA (YESAVAGE, 1983)**

- |  |                 |
|--|-----------------|
| 1.Esta satisfeito com sua vida?                              | [1] Sim [2] NÃO |
| 2.Abandonou muitas atividades e interesses?                  | [1] SIM [2] Não |
| 3.Sente que a sua vida está vazia?                           | [1] SIM [2] Não |
| 4.Sente-se frequentemente aborrecido?                        | [1] SIM [2] Não |
| 5.Esta bem disposto a maior parte do tempo?                  | [1] Sim [2] NÃO |
| 6.Tem medo que lhe suceda algo de mal?                       | [1] SIM [2] Não |
| 7.Sente-se feliz a maior parte do tempo?                     | [1] Sim [2] NÃO |
| 8.Sente-se frequentemente sem esperança?                     | [1] SIM [2] Não |
| 9.Prefere ficar em casa mais do que sair fazer coisas novas? | [1] SIM [2] Não |
| 10.Acha que tem mais problemas de memória do que a maioria?  | [1] SIM [2] Não |
| 11.Acredita que é maravilhoso estar vivo?                    | [1] Sim [2] NÃO |
| 12.Pensa que, tal como está agora, não vale para nada?       | [1] SIM [2] Não |
| 13.Pensa que a sua situação é desesperadora?                 | [1] SIM [2] Não |
| 14.Sente-se cheio de energia?                                | [1] Sim [2] NÃO |
| 15.Acha que a maioria das pessoas está melhor que você?      | [1] SIM [2] Não |

Se as respostas coincidem com a opção em maiúscula (sim, anotar um 1 ponto)

Avaliação

0 - 4: normal

5 - 10: depressão moderada

>10: depressão grave

Data: \_\_\_\_/\_\_\_\_/\_\_\_\_

**Avaliador**

## 11 - FICHA ATIVIDADE FÍSICA

### QUESTIONÁRIO INTERNACIONAL DE ATIVIDADE FÍSICA – VERSÃO CURTA



estamos interessados em saber que tipos de atividade física as pessoas fazem como parte do seu dia ... Este projeto faz parte de um grande estudo que está sendo feito em diferentes países ao redor do mundo. Suas respostas nos ajudarão a entender que tão ativos nós somos em relação à pessoas de outros países. As perguntas estão relacionadas ao tempo que você gasta fazendo atividade física na **ÚLTIMA** semana. As perguntas incluem as atividades que você faz no trabalho, para ir de um lugar a outro, por lazer, por esporte, por exercício ou como parte das suas atividades em casa ou no jardim. Suas respostas são **MUITO** importantes. Por favor responda cada questão mesmo que considere que não seja ativo. Obrigado pela sua participação!

Para responder as questões lembre que:

1. Atividades físicas **VIGOROSAS** são aquelas que precisam de um grande esforço físico e que fazem respirar **MUITO** mais forte que o normal
2. Atividades físicas **MODERADAS** são aquelas que precisam de algum esforço físico e que fazem respirar **UM POUCO** mais forte que o normal

Para responder as perguntas pense somente nas atividades que você realiza **por pelo menos 10 minutos contínuos** de cada vez:

**1a** Em quantos dias da última semana você caminhou **por pelo menos 10 minutos contínuos** em casa ou no trabalho, como forma de transporte para ir de um lugar para outro, por lazer, por prazer ou como forma de exercício? Dias \_\_\_\_\_ por **SEMANA** ( ) Nenhum

**1b** Nos dias em que você caminhou **por pelo menos 10 minutos contínuos** quanto tempo no total você gastou caminhando **por dia**? Horas: \_\_\_\_\_ Minutos: \_\_\_\_\_

**2a.** Em quantos dias da última semana, você realizou atividades **MODERADAS** **por pelo menos 10 minutos contínuos**, como por exemplo pedalar leve na bicicleta, nadar, dançar, fazer ginástica aeróbica leve, jogar vôlei recreativo, carregar pesos leves, fazer serviços domésticos na casa, no quintal ou no jardim como varrer, aspirar, cuidar do jardim, ou qualquer atividade que fez aumentar **moderadamente** sua respiração ou batimentos do coração (**POR FAVOR NÃO INCLUA CAMINHADA**) Dias \_\_\_\_\_ por **SEMANA** ( ) Nenhum

**2b.** Nos dias em que você fez essas atividades moderadas **por pelo menos 10 minutos contínuos**, quanto tempo no total você gastou fazendo essas atividades **por dia**? Horas: \_\_\_\_\_ Minutos: \_\_\_\_\_

**3a** Em quantos dias da última semana, você realizou atividades **VIGOROSAS** **por pelo menos 10 minutos contínuos**, como por exemplo, correr, fazer ginástica aeróbica, jogar futebol, pedalar rápido na bicicleta, jogar basquete, fazer serviços domésticos pesados em casa, no quintal ou cavoucar no jardim, carregar pesos elevados ou qualquer atividade que fez aumentar **MUITO** sua respiração ou batimentos do coração. Dias \_\_\_\_\_ por **SEMANA** ( ) Nenhum

**3b** Nos dias em que você fez essas atividades vigorosas **por pelo menos 10 minutos contínuos** quanto tempo no total você gastou fazendo essas atividades **por dia**? Horas: \_\_\_\_\_ Minutos: \_\_\_\_\_

Estas últimas questões são sobre o tempo que você permanece sentado todo dia no trabalho, na escola ou faculdade, em casa e durante seu tempo livre. Isto inclui o tempo sentado estudando, sentado enquanto descansa, fazendo lição de casa visitando um amigo, lendo, sentado ou deitado assistindo TV. Não inclua o tempo gasto sentando durante o transporte em ônibus, trem, metrô ou carro.

**4a.** Quanto tempo no total você gasta sentado durante um **dia de semana**?

\_\_\_\_\_ horas \_\_\_\_\_ minutos

**4b.** Quanto tempo no total você gasta sentado durante em um **dia de final de semana**?

\_\_\_\_\_ horas \_\_\_\_\_ minutos

Escore:

Classificação:

**1. MUITO ATIVO:** aquele que cumpriu as recomendações de:

- a) VIGOROSA:  $\geq 5$  dias/sem e  $\geq 30$  minutos por sessão ou
- b) VIGOROSA:  $\geq 3$  dias/sem e  $\geq 20$  minutos por sessão + MODERADA ou CAMINHADA:  $\geq 5$  dias/sem e  $\geq 30$  minutos por sessão.

**2. ATIVO:** aquele que cumpriu as recomendações de:

- a) VIGOROSA:  $\geq 3$  dias/sem e  $\geq 20$  minutos por sessão; ou
- b) MODERADA ou CAMINHADA:  $\geq 5$  dias/sem e  $\geq 30$  minutos por sessão; ou
- c) Qualquer atividade somada:  $\geq 5$  dias/sem e  $\geq 150$  minutos/sem (caminhada + moderada + vigorosa).

**3. IRREGULARMENTE ATIVO:** aquele que realiza atividade física, porém, de forma insuficiente para ser classificado como ativo pois não cumpre as recomendações quanto à freqüência ou duração. Para realizar essa classificação soma-se a freqüência e a duração dos diferentes tipos de atividades (caminhada + moderada + vigorosa).

**4. SEDENTÁRIO:** aquele que não realizou nenhuma atividade física por pelo menos 10 minutos contínuos durante a semana.

**Exemplos:**

Indivíduos	Caminhada		Moderada		Vigorosa		Classificação
	F	D	F	D	F	D	
1	-	-	-	-	-	-	Sedentário
2	4	20	1	30	-	-	Irregularmente Ativo
3	3	30	-	-	-	-	Irregularmente Ativo
4	3	20	3	20	1	30	Ativo
5	5	45	-	-	-	-	Ativo
6	3	30	3	30	3	20	Muito Ativo
7	-	-	-	-	5	30	Muito Ativo

F = Freqüência – D = Duração

Data: \_\_\_\_ / \_\_\_\_ / \_\_\_\_

Avaliador

## 12 - TIME UP AND GO

Tempo para completar \_\_\_\_\_ segundos

### Instruções:

O paciente deve levantar-se de uma cadeira de braço, sem o apoio de braços, caminhar três metros com passos seguros e confortáveis, girar 180º, retornar, sentando-se na cadeira. O tempo no qual o idoso realiza essa tarefa é cronometrado.

Recomendações necessárias na aplicação do teste:

- A altura aproximada do assento da cadeira é de 46 cm
- O paciente inicia o teste recostado
- Sempre que possível, o paciente deverá ser treinado no teste, previamente
- O tempo é cronometrado a partir do comando de partida até o paciente assentar-se novamente na cadeira
- O paciente deve usar calçado usual e, até mesmo, seu dispositivo de ajuda
- O trajeto deve ser sinalizado no chão com uma faixa crepe

### Classificação quanto à mobilidade:

[1] **Normal**– Nenhum sinal de risco de quedas;

[2] **Anormalidade leve**– Base de apoio maior ou em menor velocidade;

[3] **Anormalidade média**– hesitar, demonstrar movimentos descoordenados, velocidade insegura;

[4] **Anormalidade moderada**– Problemas ao manter sentado ou ao sentar, sendo necessária supervisão;

[5] **Anormalidade grave** – Risco claro de queda, sendo necessário suporte físico.

### Classificação quanto ao risco de queda:

[1] <13,5 segundos: baixo risco de quedas

[2] 13,5 a 20 segundos: médio risco de quedas

[3] > de 20 segundos: alto risco de quedas

Data: \_\_\_\_ / \_\_\_\_ / \_\_\_\_

Avaliador

**13 - PPS**

**Palliative Performance Scale – PPS Versão 2:**

%	Deambulação	Atividade e evidência da doença	Auto-cuidado	Ingesta	Nível da Consciência
100	Completa	Atividade normal e trabalho; sem evidência de doença	Completo	Normal	Completa
90	Completa	Atividade normal e trabalho; alguma evidência de doença	Completo	Normal	Completa
80	Completa	Atividade normal com esforço; alguma evidência de doença	Completo	Normal ou reduzida	Completa
70	Reduzida	Incapaz para o Trabalho; Doença significativa	Completo	Normal ou reduzida	Completa
60	Reduzida	Incapaz para o hobbies/trabalho doméstico. Doença significativa	Assistência ocasional.	Normal ou reduzida	Completa ou períodos de Confusão.
50	Maior parte de tempo sentado ou deitado	Incapacitado para qualquer trabalho; Doença extensa.	Assistência Considerável	Normal ou reduzida	Completa ou períodos de Confusão.
40	Maior parte do tempo acamado	Incapaz para a maioria das atividades . Doença extensa	Assistência quase completa	Normal ou reduzida	Completa ou sonolência, +/- confusão
30	Totalmente acamado	Incapaz para qualquer atividade. Doença extensa	Dependência Completa	Normal ou Reduzida	Completa ou sonolência, +/- confusão
20	Totalmente acamado	Incapaz para qualquer atividade. Doença extensa	Dependência Completa	Minima a pequenos goles	Completa ou sonolência, +/- confusão
10	Totalmente acamado	Incapaz para qualquer atividade. Doença extensa	Dependência Completa	Cuidados com a boca	Sonolência ou coma, +/- confusão
0	Morte	-	-	-	-

Escore: \_\_\_\_\_

Data: \_\_\_\_ / \_\_\_\_ / \_\_\_\_

Avaliador

### 14- INDICE DE KATZ - AVD

#### TOMAR BANHO:

- não recebe ajuda (entra e sai da banheira sozinho, se este for o modo habitual de tomar banho) (I).
- recebe ajuda para lavar apenas uma parte do corpo (como, por exemplo, as costas ou uma perna) (I).
- recebe ajuda para lavar mais de uma parte do corpo, ou não toma banho sozinho (D).

#### VESTIR-SE

- pega as roupas e veste-se completamente, sem ajuda (I)
- pega as roupas e veste-se sem ajuda, exceto para amarrar os sapatos (I)
- recebe ajuda para pegar as roupas ou vestir-se, ou permanece parcial ou completamente sem roupa (D)

#### USO DO VASO SANITÁRIO

- vai ao banheiro ou local equivalente, limpa-se e ajeita as roupas sem ajuda (pode usar objetos para apoio como bengala, andador ou cadeira de rodas e pode usar comadre ou urinol à noite, esvaziando-o de manhã) (I)
- recebe ajuda para ir ao banheiro ou local equivalente, ou para limpar-se, ou para ajeitar as roupas após evacuação ou micção, ou para usar a comadre ou urinol à noite (D)
- não vai ao banheiro ou equivalente para eliminações fisiológicas (D)

#### TRANSFERÊNCIA

- deita-se e sai da cama, senta-se e levanta-se da cadeira sem ajuda (pode estar usando objeto para apoio, como bengala ou andador) (I)
- deita-se e sai da cama e/ou senta-se e levanta-se da cadeira com ajuda (D)
- não sai da cama (D)

#### CONTINÊNCIA

- controla inteiramente a micção e a evacuação (I)
- tem “acidentes” ocasionais (D)
- necessita de ajuda para manter o controle da micção e evacuação; usa cateter ou é incontinente(D)

#### ALIMENTAÇÃO

- alimenta-se sem ajuda (I)
- alimenta-se sozinho, mas recebe ajuda para cortar carne ou passar manteiga no pão (I)
- recebe ajuda para alimentar-se, ou é alimentado parcialmente ou completamente pelo uso de cateteres. (D)

#### INTERPRETAÇÃO:

- 0: independente em todas as seis funções;
- 1: independente em cinco funções e dependente em uma função;
- 2: independente em quatro funções e dependente em duas;
- 3: independente em três funções e dependente em três;
- 4: independente em duas funções e dependente em quatro;
- 5: independente em uma função e dependente em cinco funções;
- 6: dependente em todas as seis funções.

Escore : \_\_\_\_\_

Data: \_\_\_\_/\_\_\_\_/\_\_\_\_\_

Avaliador

Local etiqueta ou identificação do/a paciente

**15 – QLQ-C30**

BRAZILIAN



**EORTC QLQ-C30 (versão 3.0.)**

Nós estamos interessados em alguns dados sobre você e sua saúde. Responda, por favor, a todas as perguntas fazendo um círculo no número que melhor se aplica a você. Não há respostas certas ou erradas. A informação que você fornecer permanecerá estritamente confidencial.

Por favor, preencha suas iniciais:

Sua data de nascimento (dia, mês, ano):

Data de hoje (dia, mês, ano):

31							
----	--	--	--	--	--	--	--

	<b>Não</b>	<b>Pouco</b>	<b>Modera- damente</b>	<b>Muito</b>
1. Você tem qualquer dificuldade quando faz grandes esforços, por exemplo carregar uma bolsa de compras pesada ou uma mala?	1	2	3	4
2. Você tem qualquer dificuldade quando faz uma <u>longa</u> caminhada?	1	2	3	4
3. Você tem qualquer dificuldade quando faz uma <u>curta</u> caminhada fora de casa?	1	2	3	4
4. Você tem que ficar numa cama ou na cadeira durante o dia?	1	2	3	4
5. Você precisa de ajuda para se alimentar, se vestir, se lavar ou usar o banheiro?	1	2	3	4

**Durante a última semana:**

	<b>Não</b>	<b>Pouco</b>	<b>Modera- damente</b>	<b>Muito</b>
6. Tem sido difícil fazer suas atividades diárias?	1	2	3	4
7. Tem sido difícil ter atividades de divertimento ou lazer?	1	2	3	4
8. Você teve falta de ar?	1	2	3	4
9. Você tem tido dor?	1	2	3	4
10. Você precisou repousar?	1	2	3	4
11. Você tem tido problemas para dormir?	1	2	3	4
12. Você tem se sentido fraco/a?	1	2	3	4
13. Você tem tido falta de apetite?	1	2	3	4
14. Você tem se sentido enjoado/a?	1	2	3	4
15. Você tem vomitado?	1	2	3	4

Por favor, passe à pagina seguinte

Local etiqueta ou identificação do/a paciente

**Durante a Última semana:**

	<b>Não</b>	<b>Pouco</b>	<b>Modera- damente</b>	<b>Muito</b>
16. Você tem tido prisão de ventre?	1	2	3	4
17. Você tem tido diarréia?	1	2	3	4
18. Você esteve cansado/a?	1	2	3	4
19. A dor interferiu em suas atividades diárias?	1	2	3	4
20. Você tem tido dificuldade para se concentrar em coisas, como ler jornal ou ver televisão?	1	2	3	4
21. Você se sentiu nervoso/a?	1	2	3	4
22. Você esteve preocupado/a?	1	2	3	4
23. Você se sentiu irritado/a facilmente?	1	2	3	4
24. Você se sentiu deprimido/a?	1	2	3	4
25. Você tem tido dificuldade de se lembrar das coisas?	1	2	3	4
26. A sua condição física ou o tratamento médico tem interferido em sua vida <u>familiar</u> ?	1	2	3	4
27. A sua condição física ou o tratamento médico tem interferido em suas atividades <u>sociais</u> ?	1	2	3	4
28. A sua condição física ou o tratamento médico tem lhe trazido dificuldades financeiras?	1	2	3	4

**Para as seguintes perguntas, por favor, faça um círculo em volta do número entre 1 e 7 que melhor se aplica a você.**

29. Como você classificaria a sua saúde em geral, durante a última semana?

1	2	3	4	5	6	7
Péssima						Ótima

30. Como você classificaria a sua qualidade de vida geral, durante a última semana?

1	2	3	4	5	6	7
Péssima						Ótima

© Copyright 1995, 1996 EORTC Study Group on Quality of Life. Todos os direitos reservados. Versão 3.0

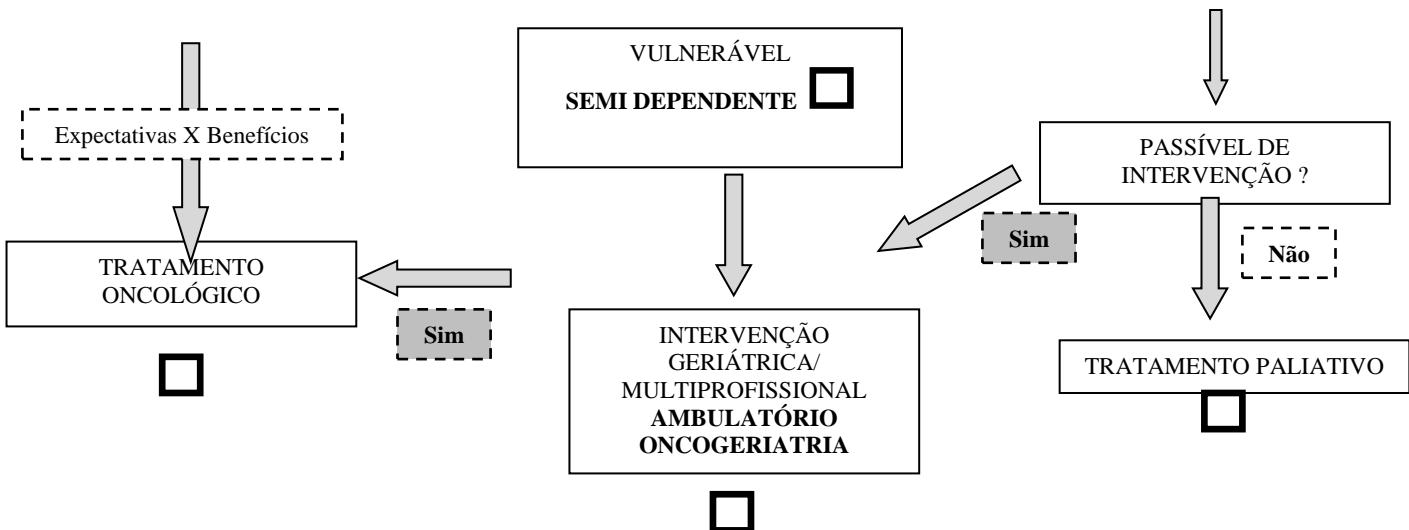
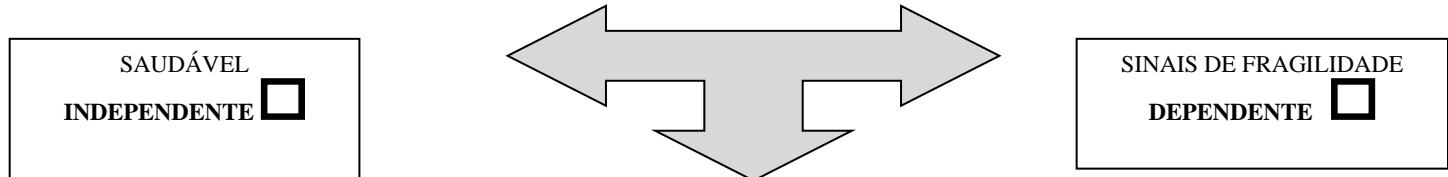
Data: \_\_\_\_ / \_\_\_\_ / \_\_\_\_

Avaliador

Local etiqueta ou identificação do/a paciente

**16 – DEFINIÇÃO DA DEPENDÊNCIA**  
**AVALIAÇÃO GERIÁTRICA AMPLA**

FUNCIONALIDADE  
 COMORBIDADES  
 CONDIÇÕES SOCIODEMOGRÁFICAS  
 SIDROMES GERIÁTRICAS  
 POLIFARMÁCIA  
 NUTRIÇÃO  
 QUALIDADE DE VIDA



**ENCAMINHAMENTO:**

- Ambulatório de Oncogeriatría
- Ambulatório de Cuidados Paliativos

Data \_\_\_\_ / \_\_\_\_ / \_\_\_\_

Avaliadores: \_\_\_\_\_

## 17 - RESULTADOS DE EXAMES E FECHAMENTO DO ESTADIAMENTO

Data coleta	Exame	Resultado
	Hemoglobina	maior g/dl
		menor g/dl
	Contagem de leucócitos	maior cel/m <sup>3</sup>
		menor cel/m <sup>3</sup>
	Nº de granulócitos	maior cel/m <sup>3</sup>
		menor cel/m <sup>3</sup>
	Plaquetas	maior cel/m <sup>3</sup>
		menor cel/m <sup>3</sup>
	Creatinina	maior mg/dl
		menor mg/dl
	Perfil imunológico	
	Expressão toll like 2, 4 e 9	

Outros resultados de exames relevantes:

Data realização	EXAME	Resultado

Fechamento do estadiamento do tumor:

Tumor de \_\_\_\_\_ CID 10: C \_\_\_\_\_

ESTADIAMENTO resumo :

[1] 0 [2] I [3] II A [4] II B [5] III A [6] III B [7] IV [9] sem informação

Metástases: [1] Sim [2] Não

Sítios: [1] óssea [2] pulmonar [3] SNC [4] hepática [5]

outras: \_\_\_\_\_

Data: \_\_\_\_ / \_\_\_\_ / \_\_\_\_

Avaliador

Ficha de seguimento <i>Se sim para algum dado verificar dados prontuário e complementar</i>	<b>Seguimento 1 30 dias</b>	<b>Seguimento 2 60 dias</b>	<b>Seguimento 3 90 dias</b>
Data programada	/ /	/ /	/ /
Data do telefonema ou visita	/ /	/ /	/ /
Uso de estimulante de colônia de neutrófilos	[1]Sim / / [2]Não	[1]Sim / / [2]Não	[1]Sim / / [2]Não
• Tempo/duração			
Antibioticoprofilaxia <i>Especifique início e término/finalidade</i>	[1]Sim / / [2]Não	[1]Sim / / [2]Não	[1]Sim / / [2]Não
• Qual:			
Infecção desde último contato			
• Não			
• Sim - <i>Preencher ficha específica para cada infecção</i>			
• Data	/ /	/ /	/ /
Topografia da infecção - especifique:			
Atendimento de urgência ou internamento	[1]Sim / / [2]Não	[1]Sim / / [2]Não	[1]Sim / / [2]Não
• Sim – quantas vezes <i>Preencher ficha específica para cada atendimento ou internamento</i>			
• Local			
Procedimentos oncológicos:			
Nenhum			
Cirurgia	[1]Sim / / [2]Não	[1]Sim / / [2]Não	[1]Sim / / [2]Não
• Qual:			
Topografia:			
Hormonioterapia	[1]Sim / / [2]Não	[1]Sim / / [2]Não	[1]Sim / / [2]Não
• Tipo (droga/classe)			
• Finalidade*:			
Quimioterapia	[1]Sim / / [2]Não	[1]Sim / / [2]Não	[1]Sim / / [2]Não
• *finalidade:			
• Tipo :			
• Mono (1 droga)			
• Poli ( $\geq$ 2 drogas)			
• Esquema (drogas):			
Anticorpo monoclonal:	[1]Sim / / [2]Não	[1]Sim / / [2]Não	[1]Sim / / [2]Não
• *finalidade			
Radioterapia	[1]Sim / / [2]Não	[1]Sim / / [2]Não	[1]Sim / / [2]Não
• Topografia			
• Dose total realizada			
Corticoide	[1]Sim / / [2]Não	[1]Sim / / [2]Não	[1]Sim / / [2]Não
Tipo de corticoide/ dose			
• Duração			
Imunoterapia	[1]Sim / / [2]Não	[1]Sim / / [2]Não	[1]Sim / / [2]Não
• Esquema :			
Transplante medula óssea	[1]Sim / / [2]Não	[1]Sim / / [2]Não	[1]Sim / / [2]Não
• Tipo: autólogo ou alogênico			
Outro especificar:			
Neutropenia febril	[1]Sim / / [2]Não	[1]Sim / / [2]Não	[1]Sim / / [2]Não

\*Finalidade : C -curativa/exclusiva , P-paliativa, N-neoadjuvante, A- adjuvante

Ficha de seguimento <i>Se sim para algum dado verificar dados prontuário e complementar</i>	<b>Seguimento 4 120 dias</b>	<b>Seguimento 5 150 dias</b>	<b>Seguimento 6 180 dias</b>
Data programada	/ /	/ /	/ /
Data do telefonema ou visita	/ /	/ /	/ /
Uso de estimulante de colônia de neutrófilos	[1]Sim / / [2]Não	[1]Sim / / [2]Não	[1]Sim / / [2]Não
• Tempo/duração			
Antibioticoprofilaxia <i>Especifique início</i>	[1]Sim / / [2]Não	[1]Sim / / [2]Não	[1]Sim / / [2]Não
• Qual:			
Infecção desde último contato			
• Não			
• Sim - <i>Preencher ficha específica para cada infecção</i>			
• Data	/ /	/ /	/ /
Topografia da infecção - especifique:			
Atendimento de urgência ou internamento	[1]Sim / / [2]Não	[1]Sim / / [2]Não	[1]Sim / / [2]Não
• Sim – quantas vezes <i>Preencher ficha específica para cada atendimento ou internamento</i>			
• Local			
Procedimentos oncológicos:			
Nenhum			
Cirurgia	[1]Sim / / [2]Não	[1]Sim / / [2]Não	[1]Sim / / [2]Não
• Qual:			
Topografia:			
Hormonioterapia	[1]Sim / / [2]Não	[1]Sim / / [2]Não	[1]Sim / / [2]Não
• Tipo (droga/classe)			
• Finalidade*:			
Quimioterapia	[1]Sim / / [2]Não	[1]Sim / / [2]Não	[1]Sim / / [2]Não
• *finalidade:			
• Tipo :			
• Mono (1 droga)			
• Poli ( $\geq$ 2drogas)			
• Esquema (drogas):			
Anticorpo monoclonal:	[1]Sim / / [2]Não	[1]Sim / / [2]Não	[1]Sim / / [2]Não
• *finalidade			
Radioterapia	[1]Sim / / [2]Não	[1]Sim / / [2]Não	[1]Sim / / [2]Não
• Topografia			
• Dose total realizada			
Corticoide	[1]Sim / / [2]Não	[1]Sim / / [2]Não	[1]Sim / / [2]Não
Tipo de corticoide/ dose			
• Duração			
Imunoterapia	[1]Sim / / [2]Não	[1]Sim / / [2]Não	[1]Sim / / [2]Não
• Esquema :			
Transplante medula óssea	[1]Sim / / [2]Não	[1]Sim / / [2]Não	[1]Sim / / [2]Não
• Tipo: autólogo ou alogênico			
Outro especificar:			
Neutropenia febril	[1]Sim / / [2]Não	[1]Sim / / [2]Não	[1]Sim / / [2]Não

\*Finalidade : C -curativa/exclusiva , P-paliativa, N-neoadjuvante, A- adjuvante

Local etiqueta ou identificação do/a paciente

## 24 – FICHA DE INTERNAMENTO /SPA

### Ficha de Internamento ou SPA Nº

Data de admissão \_\_\_\_/\_\_\_\_/\_\_\_\_

Local de internação

1 - SPA

5 – Cuidados Paliativos

2 – Enfermaria Oncologia

6 – UTI

3 – Enfermaria Hematologia

7 – Outros – Especifique

4 – Enfermaria Cirúrgica

Motivo da internação: ( ) Clínico ( ) Cirúrgico \_\_\_\_\_

Neutropenia febril: [1]Sim [2] Não Data: \_\_\_\_/\_\_\_\_/\_\_\_\_

**Colonização por microrganismo multiresistente :** [1]Sim [2] Não Data: / /

**Microrganismo colonização:** \_\_\_\_\_

IrAS : [1]Sim [2] Não

Data: \_\_\_\_/\_\_\_\_/\_\_\_\_ (preencher ficha específica para cada infecção)

Gravidade	0. Sem sepse	[1]Sim	[2] Não
	1. Sepse	[1]Sim	[2] Não
	2. Sepse grave	[1]Sim	[2] Não
	3. Choque séptico	[1]Sim	[2] Não

Disfunção de múltiplos órgãos e sistemas (DMOS): [1]Sim [2] Não

**Decisão para limitação da terapêutica:** [1]Sim [2] Não Data: \_\_\_\_/\_\_\_\_/\_\_\_\_

Observações:

---

---

---

Data da alta: / /

Local etiqueta ou identificação do/a paciente

**PROCEDIMENTOS REALIZADOS DURANTE A INTERNAÇÃO**

	Não	Sim	Data início	Data fim
<b>Cirurgia</b>				
○ Qual:				
Radioterapia				
Quimioterapia				
Aminas vasoativas				
Transfusão hemoderivados				
○ Sangue				
○ Plasma				
○ Albumina				
Antimicrobianos :				
Diálise peritoneal				
Hemodiálise				
Sonda gástrica				
<b>Sonda vesical</b>				
Nebulização				
Ventilação não invasiva				
<b>Intubação</b>				
Traqueostomia				
Drenagem pleural				
<b>Acesso venoso central</b>				
○ Punção venosa central				
○ Totalmente implantado				
○ PICC – inserção periférica				
Outro – especificar:				

Colar etiqueta

## 25 - FICHA DE NOTIFICAÇÃO IRAS

IRAS	Data	Topografia
1 <sup>a</sup>		
2 <sup>a</sup>		
3 <sup>a</sup>		

### Topografia da IrAS

- |   |  |
|---|--|
| 1-Pneumonia   | 7- Infecção do trato urinário                              |
| 2- Pneumonia associada ventilador                                     | 8 -Infecção do trato urinário associado ao cateter vesical |
| 3-Infecção de sítio cirúrgico   | 9-Osteoarticular   |
| 4-Infecção da corrente sanguínea (clinica)                            | 10- Infecção local do cateter                              |
| 5- Infecção da corrente sanguínea com confirmação bacteriológica      | 8- Olhos, ouvidos, nariz e garganta                        |
| 6- Infecção da corrente sanguínea associada ao cateter venoso central | 12-Outro   |

### Controle microbiológico – bacterioscopia (GRAM) e culturas

ESPÉCIME	Data coleta	Resultado	Microrganismo isolado
		Neg ( ) Pos ( )	
		Neg ( ) Pos ( )	
		Neg ( ) Pos ( )	
		Neg ( ) Pos ( )	
		Neg ( ) Pos ( )	
		Neg ( ) Pos ( )	
		Neg ( ) Pos ( )	
		Neg ( ) Pos ( )	
		Neg ( ) Pos ( )	
		Neg ( ) Pos ( )	
		Neg ( ) Pos ( )	
		Neg ( ) Pos ( )	

### Evolução:

**SE MICRORGANISMO ISOLADO  
RECUPERAR O RESULTADO DO ANTBIOGRAMA  
ANEXAR O RESULTADO**