Utilização de aprendizagem de máquina na classificação de risco de pacientes oncológicos

Tiago Beltrão Lacerda

Orientadora: Ana Paula Cavalcanti Furtado, D.Sc (CESAR.School / UFRPE)
Coorientadora: Dra. Jurema Telles de Oliveira Lima, D.Sc (IMIP)

Mestrado profissional em engenharia de software
Recife - 25/06/2020

Sumário

Contexto

Big data

Necessita de novos paradigmas para ser explorado (BEYER; LANEY, 2012)

Aprendizagem de máquina

Derivada da Inteligência Artificial, a AM é a chave para medicina de precisão (KOROU et al, 2015)

Medicina de Precisão

Individualizar o tratamento para todos (CHIAVEGATTO, 2015)

Câncer

Problema de saúde pública mundial, afetando principalmente idosos. Será principal causa de morte no mundo (BRAY et al., 2018)

AGA

Avaliação Geriátrica Ampla é necessária pois apenas idade não é capaz de descrever o estado de saúde de uma pessoa (WILDIERS et al., 2014)

Motivação

- AGA não foi desenvolvida especificamente para o tratamento oncológico (RAMJAUN et al., 2013)
  - Tempo de aplicação da AGA: ~45min
  - Total de 107 variáveis
  - Precisa de equipe multidisciplinar
  - ...por isso não é adotada em larga escala
  - Mesmo os centros que a aplicam, tomam decisões com base nas grandes médias

Problema

?

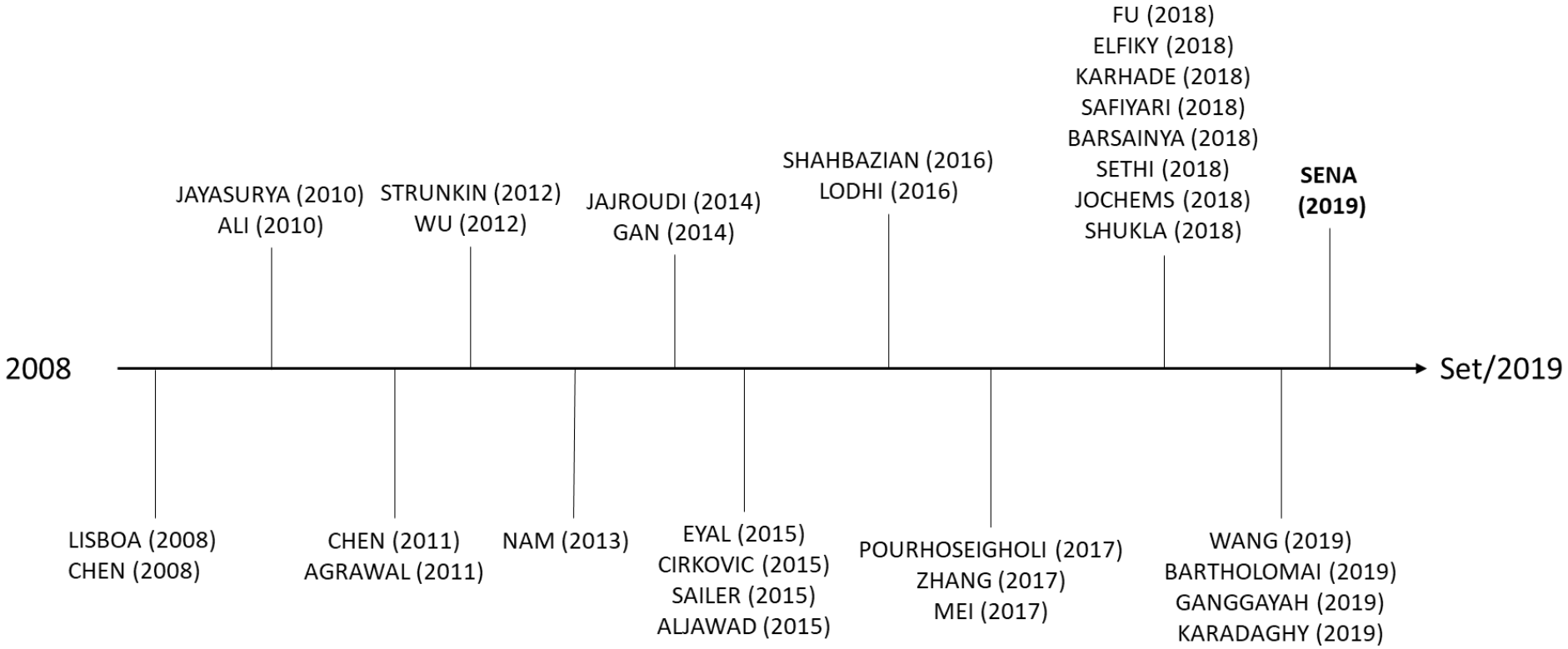
Como classificar os pacientes oncológicos em alto ou baixo risco de óbito em 6 meses utilizando as variáveis presentes na AGA?

Objetivos

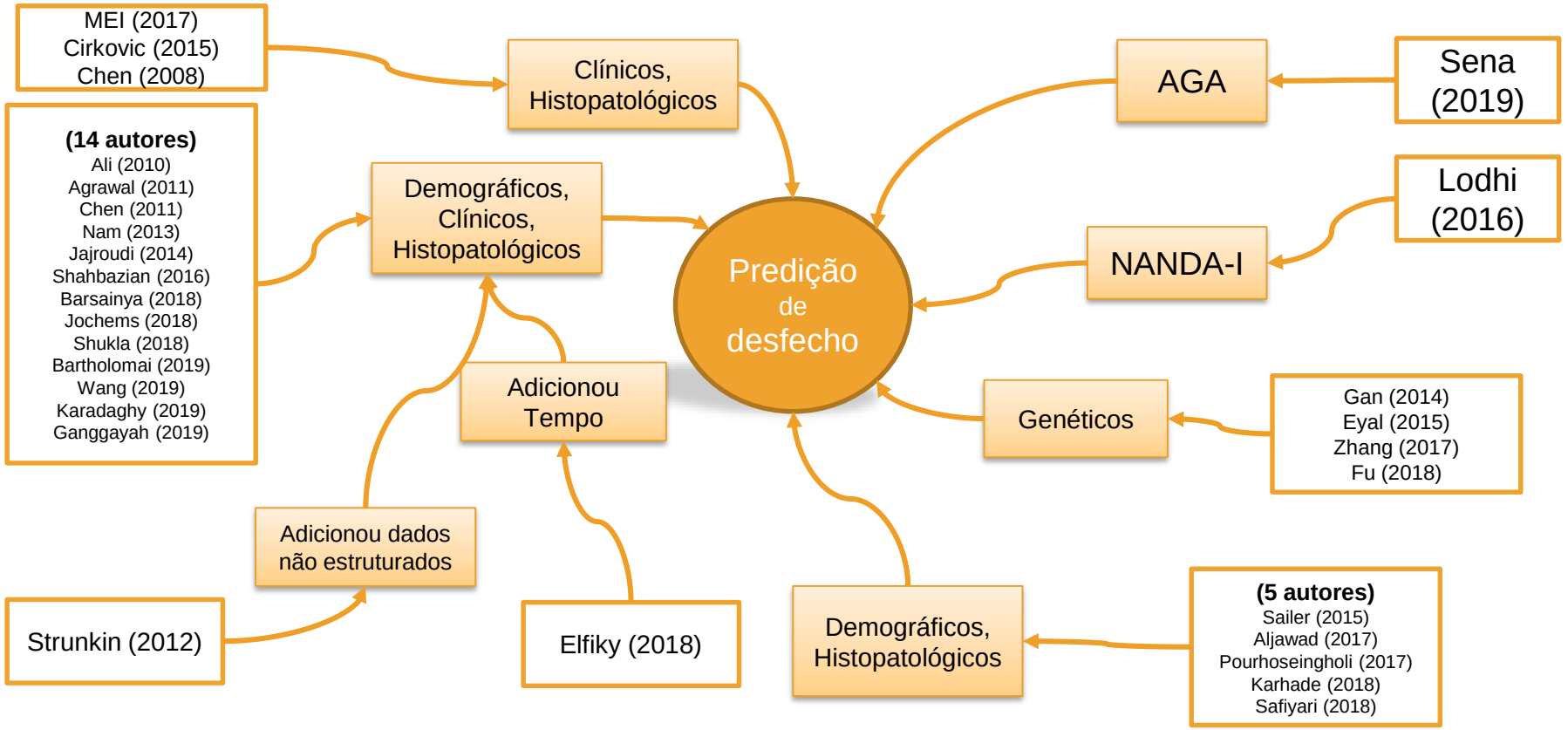
- Desenvolver um modelo de AM capaz de classificar os pacientes idosos em tratamento oncológico em alto e baixo risco de morte em 6 meses

- Determinar o conjunto mínimo de variáveis
- Avaliar qual algoritmo apresenta melhor desempenho
- Otimizar a performance do algoritmo selecionado
- Priorizar a identificação dos pacientes de alto risco

Trabalhos relacionados



Trabalhos relacionados



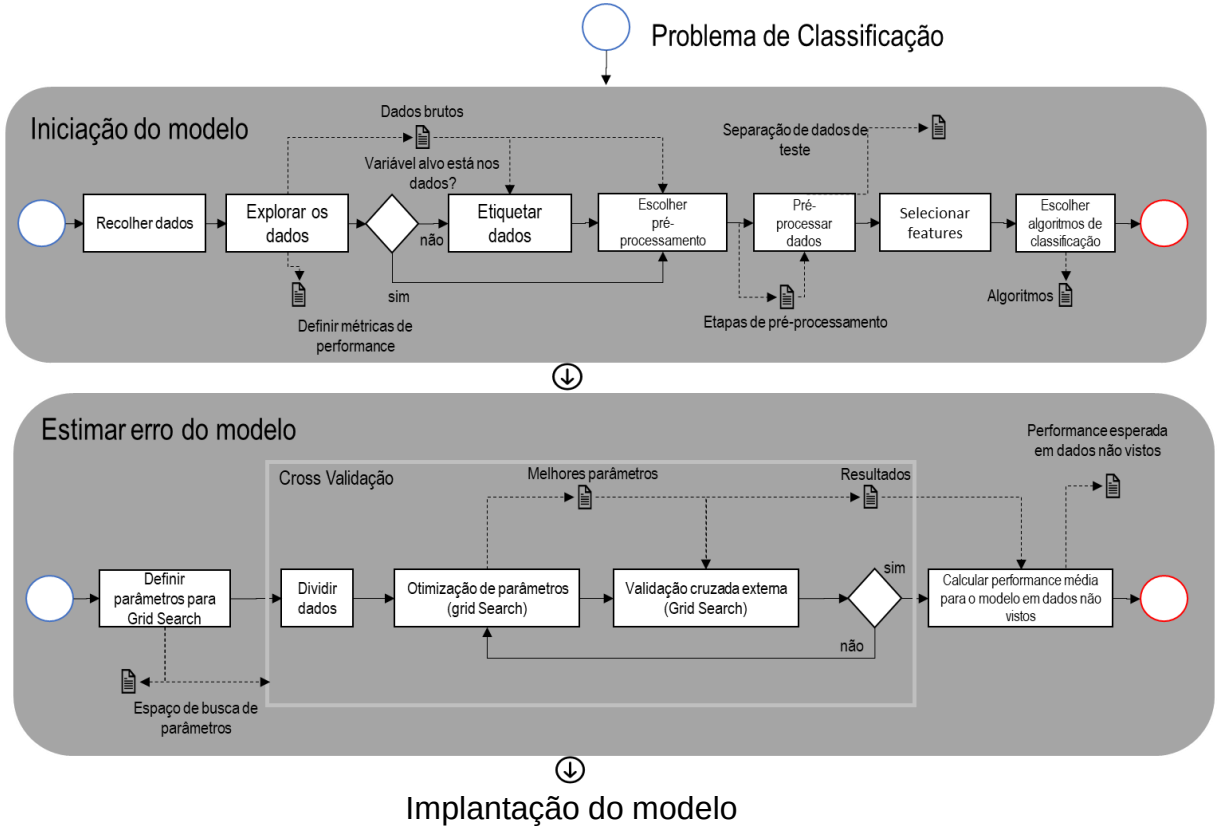
Trabalhos relacionados

- Base de dados com 608 Pacientes x 1600
- Algoritmos sem otimização x otimização
- Variáveis condensadas em escores x todas as variáveis
- AUC-ROC (0,833) x AUC-PR
- Normalização min-max x z-score

Metodologia

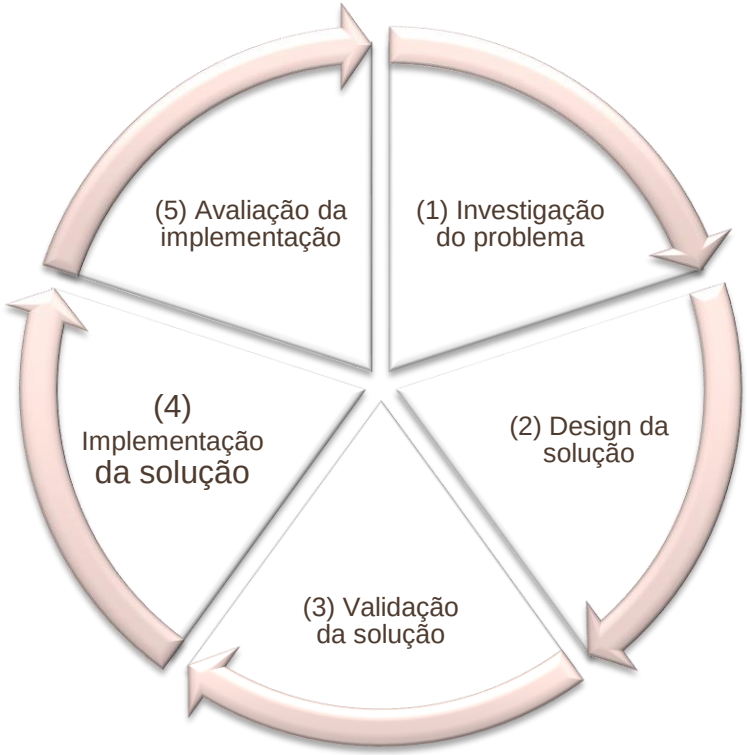
Framework para desenvolvimento de modelos de AM para problemas de classificação

- Modelo desenvolvido por Hirt, Niklas e Satzger (2017), voltado para problemas de classificação.



Design Science Research

- DSR é uma 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 (WIERINGA, 2014)
- A metodologia do Design Science é um paradigma para a condução e comunicação de pesquisa aplicada, como é o caso da engenharia de Software (Engström et al., 2020)



Ciclo de engenharia. Ciclo de design (1 a 3).

Análise e pré-processamento

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

Análise e pré-processamento dos dados

Problema de classificação

- Variável de saída assume dois valores
  - Classe positiva(1): Óbito em menos de 6 meses
  - Classe negativa(0): Vida
- Base desbalanceada, com 15,4% de óbitos em 6 meses

Base de dados

- 1600 registros
- 136 variáveis, 107 da AGA
- 103 variáveis categóricas e 33 numéricas
- 54,9% homens e 45,1% mulheres
- Tipos de câncer
  - 32,2% próstata
  - 14,9% mama

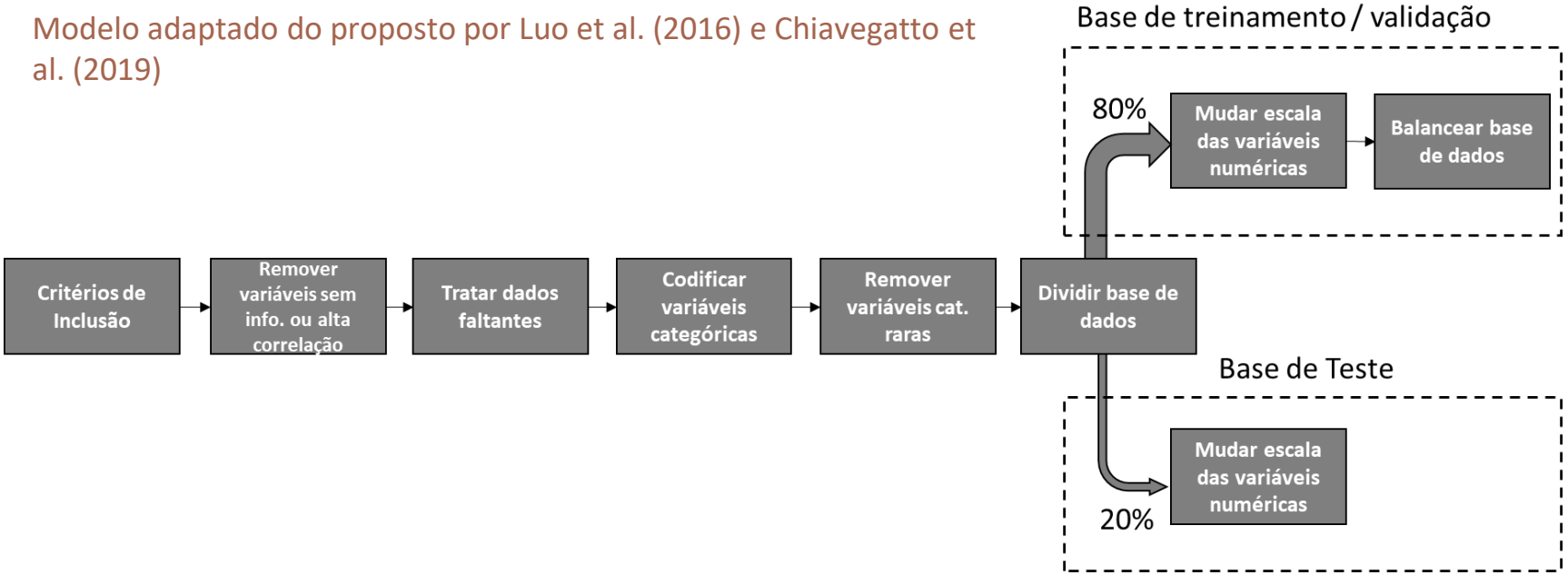
Métricas utilizadas

Utilizou-se as métricas voltadas para a classe positiva, de forma a melhorar a identificação dos mesmos

- AUC-*Precision-Recall (decisória)*
- *Recall, f1-score, AUC-ROC (informativas)*

Análise e pré-processamento dos dados

Modelo adaptado do proposto por Luo et al. (2016) e Chiavegatto et al. (2019)

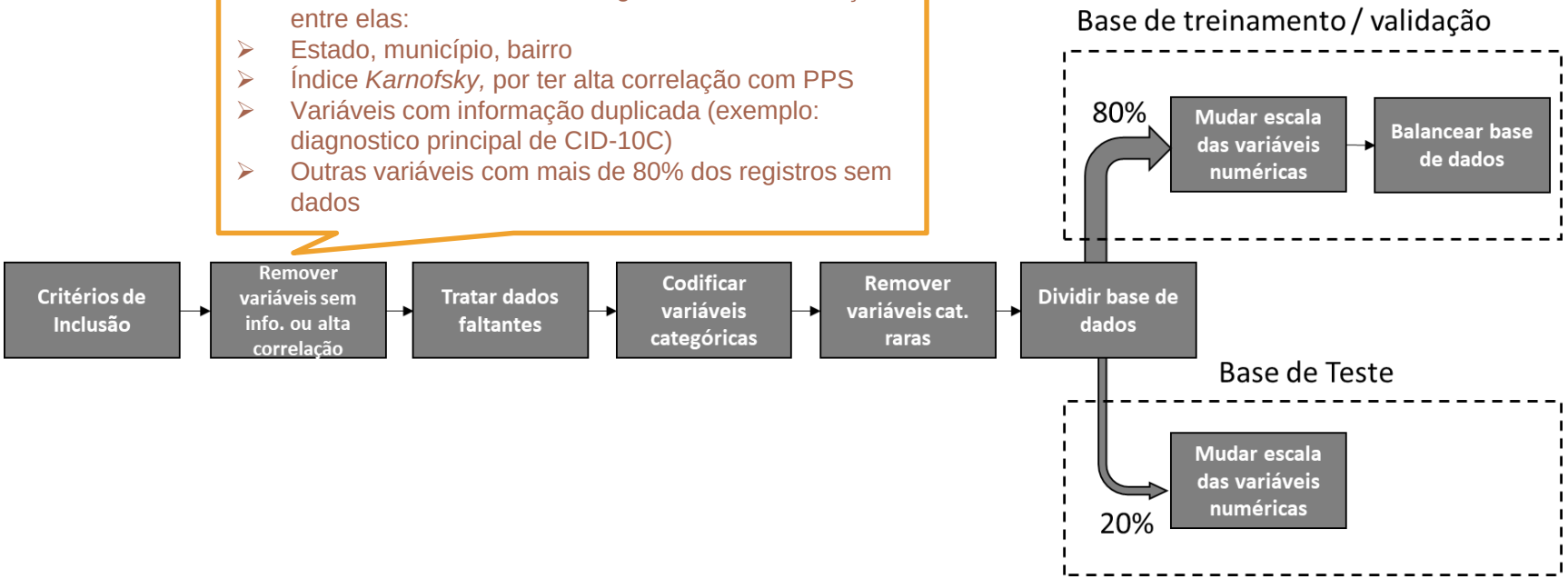


Análise e pré-processamento dos dados

- Estar na base do IMIP
- Ter completado 6 meses de acompanhamento ou falecido antes.

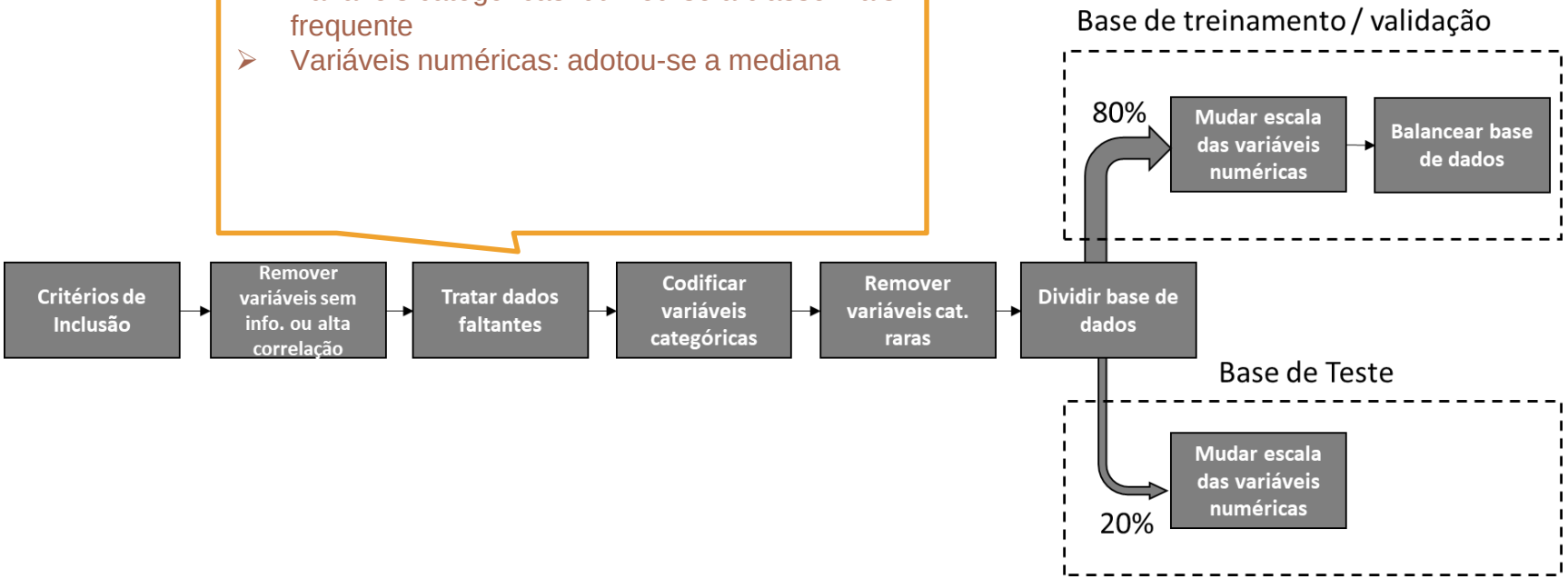
Análise e pré-processamento dos dados

- Removidas 12 variáveis categóricas sem informação entre elas:
- Estado, município, bairro
- Índice *Karnofsky,* por ter alta correlação com PPS
- Variáveis com informação duplicada (exemplo: diagnostico principal de CID-10C)
- Outras variáveis com mais de 80% dos registros sem dados



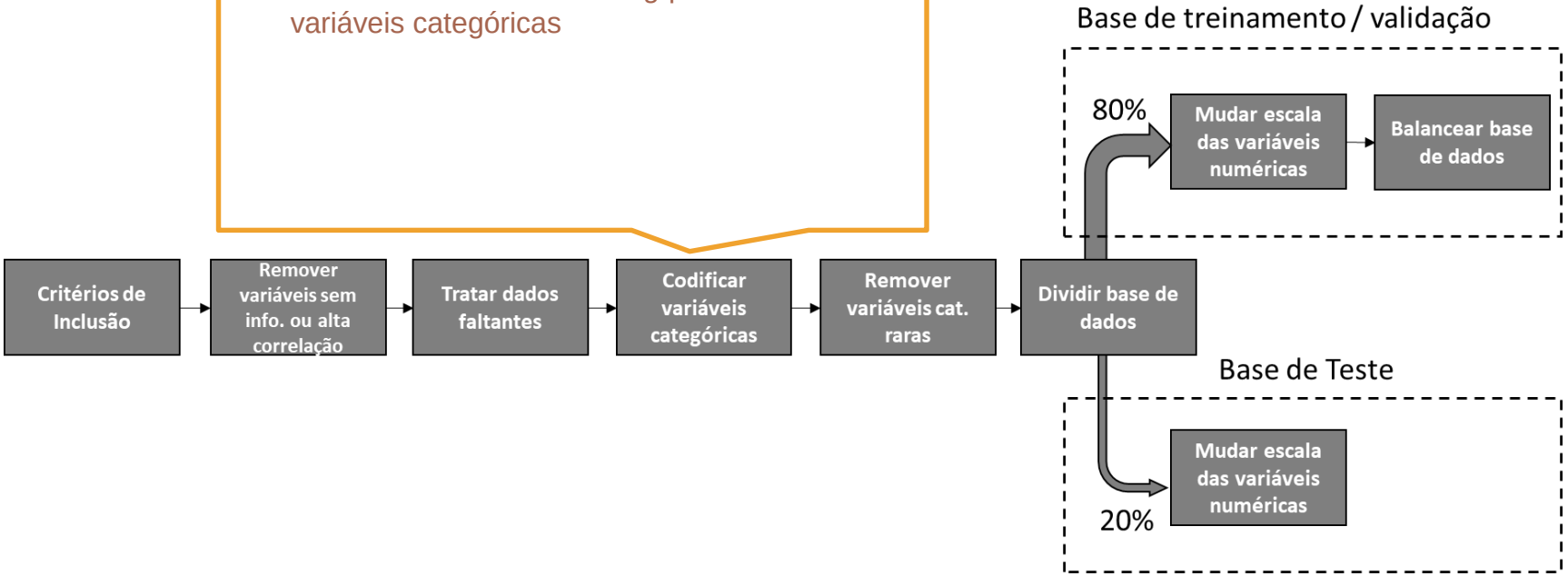
Análise e pré-processamento dos dados

- Variáveis categóricas: utilizou-se a classe mais frequente
- Variáveis numéricas: adotou-se a mediana



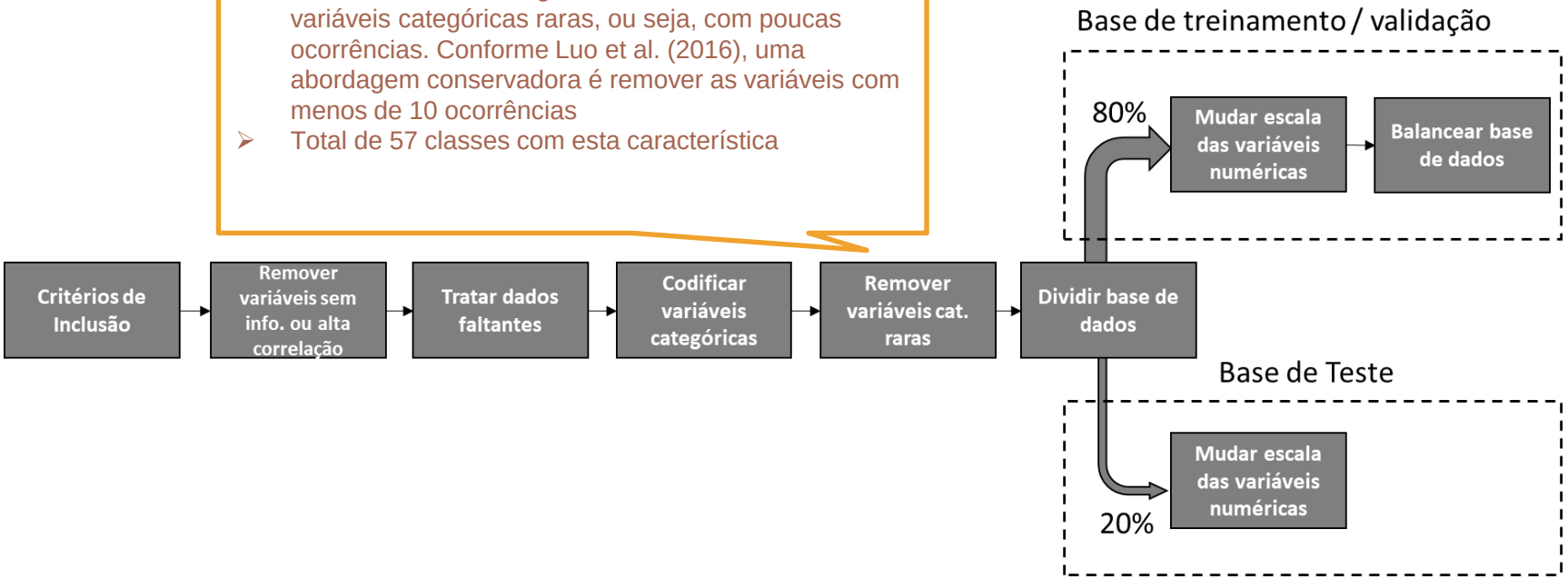
Análise e pré-processamento dos dados

- Utilizou-se *one-hot Encoding* para codificar as variáveis categóricas



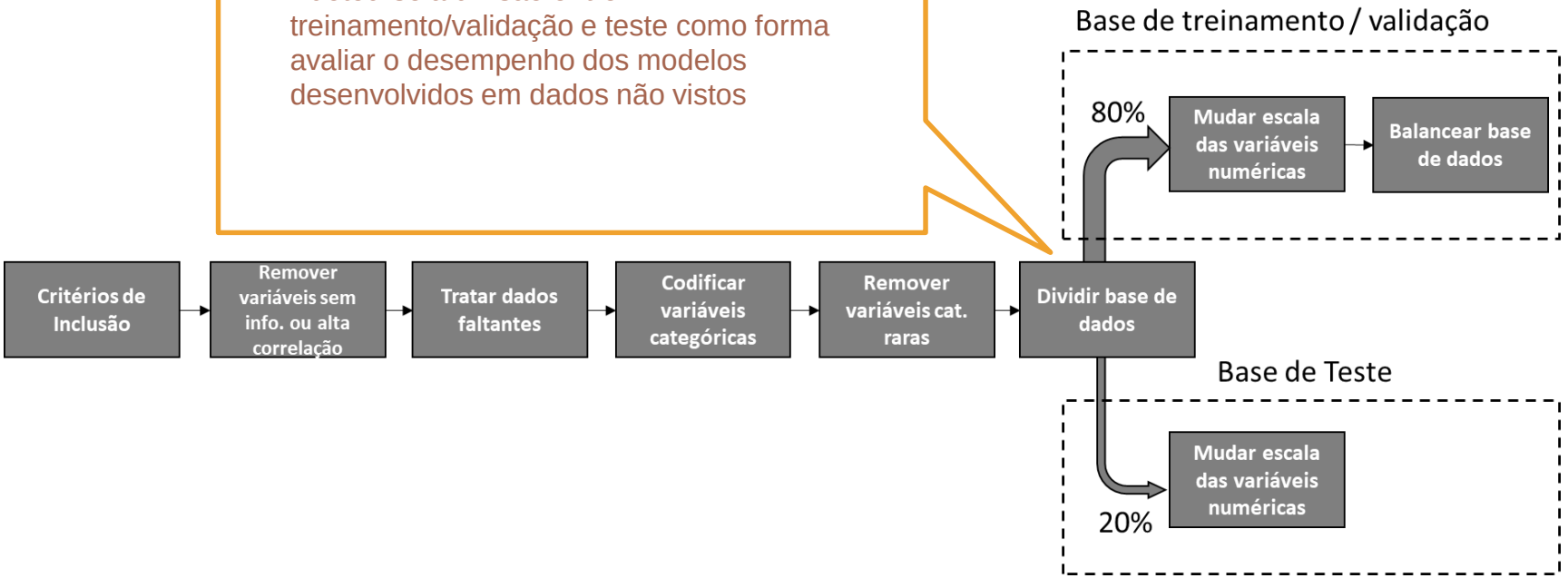
Análise e pré-processamento dos dados

- Para evitar o *overfitting*, removeu-se do *dataset* as variáveis categóricas raras, ou seja, com poucas ocorrências. Conforme Luo et al. (2016), uma abordagem conservadora é remover as variáveis com menos de 10 ocorrências
- Total de 57 classes com esta característica



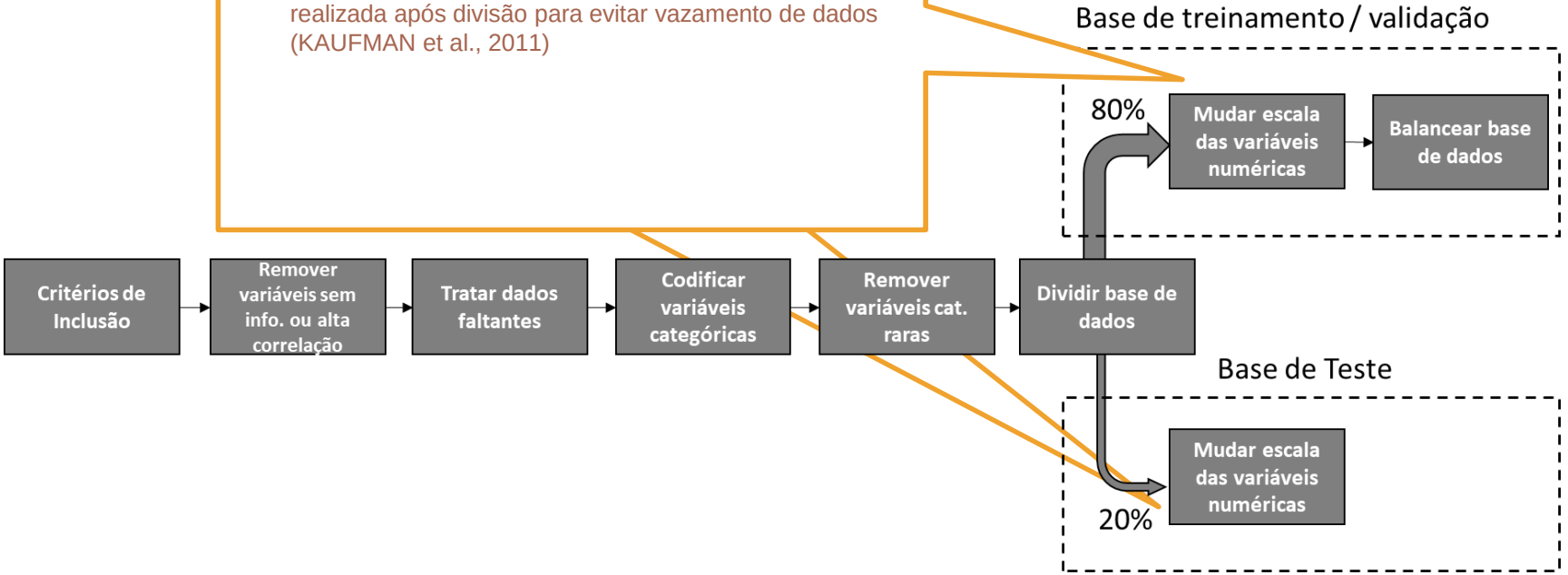
Análise e pré-processamento dos dados

- Adotou-se a divisão entre treinamento/validação e teste como forma avaliar o desempenho dos modelos desenvolvidos em dados não vistos



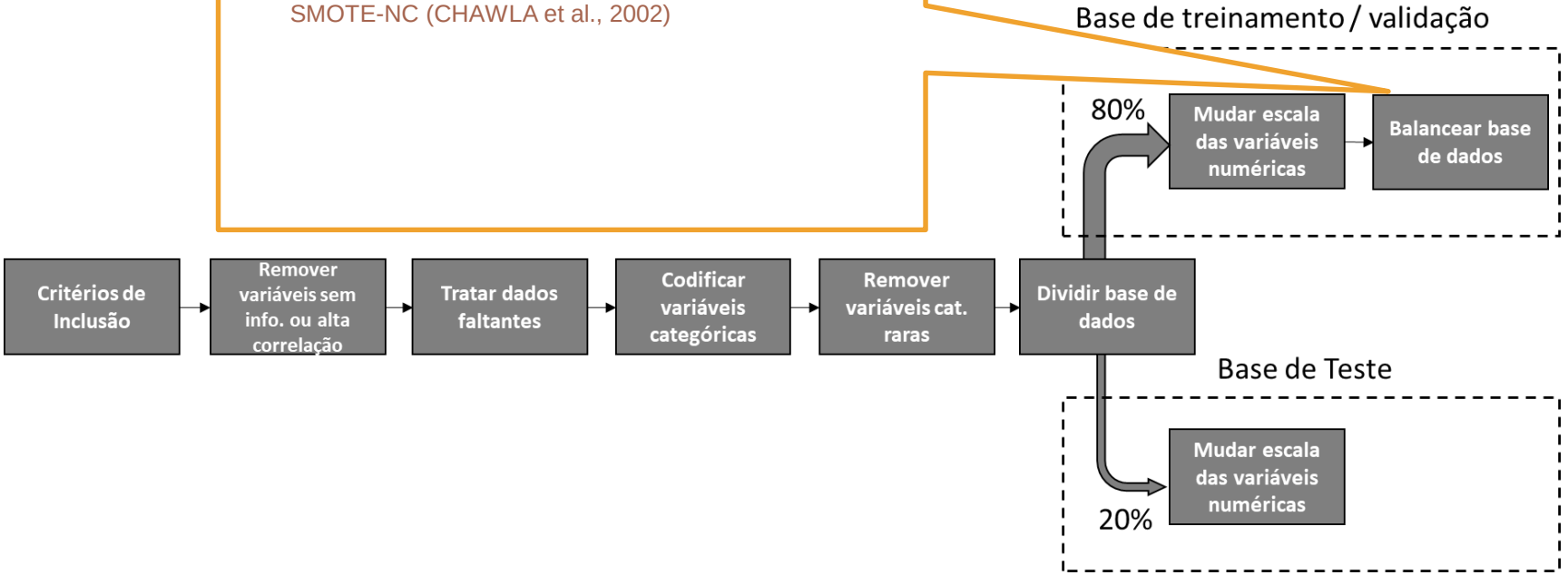
Análise e pré-processamento dos dados

- Mudança de escala (*z-score*) nas variáveis numéricas realizada após divisão para evitar vazamento de dados (KAUFMAN et al., 2011)

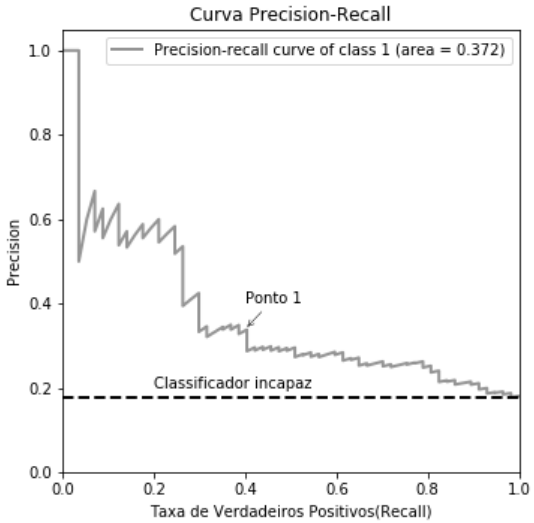


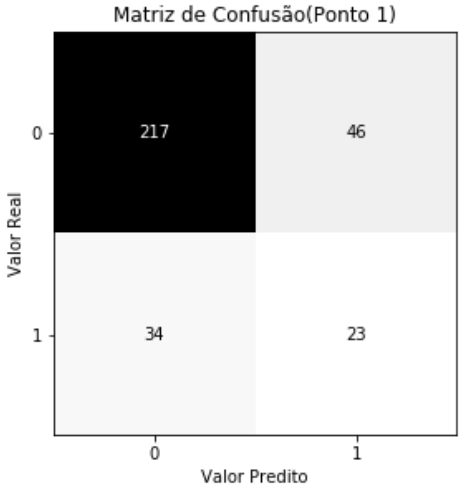
Análise e pré-processamento dos dados

- Base de treinamento/validação balanceada por meio do SMOTE-NC (CHAWLA et al., 2002)



Modelo de Referência





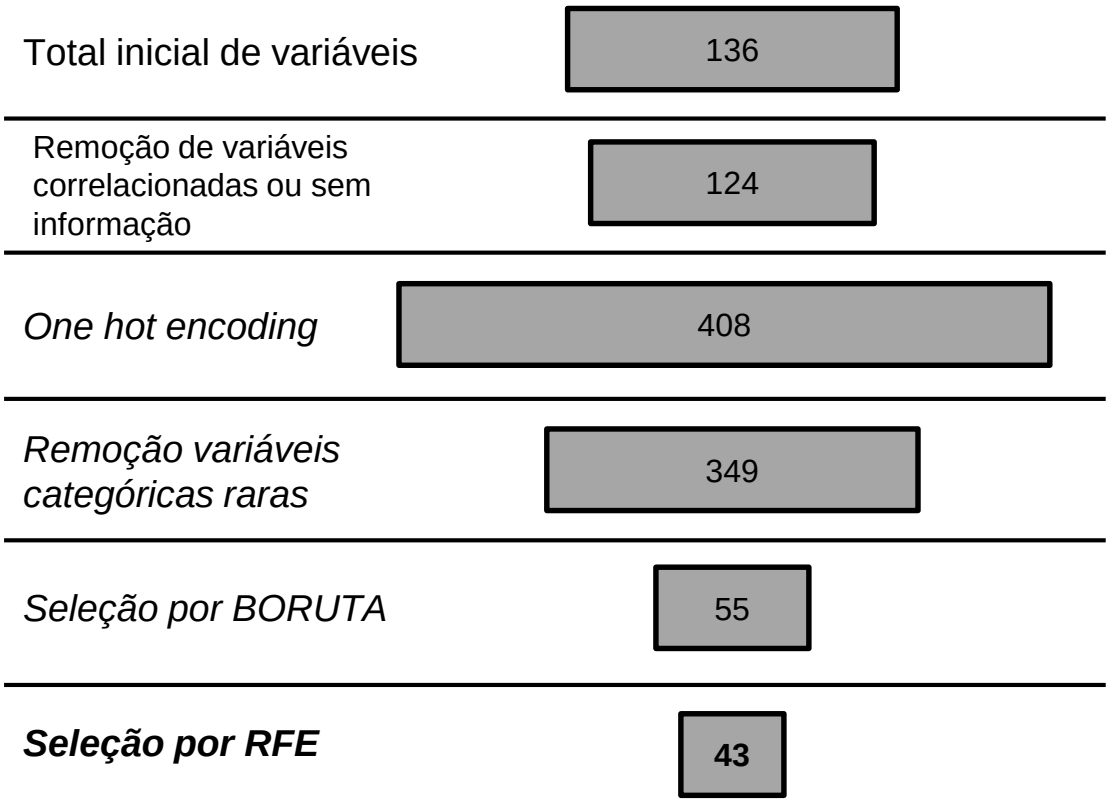
Modelo de referência utilizando Regressão Logística e todos as *features* do *dataset*

Algoritmo	AUC PR	Recall	AUC ROC
Regressão Logística (Referência)	**0,372**	**0,404**	**0,683**

DSR: Ciclo 1

Quais variáveis mais importantes? Quais as minimamente necessárias? Qual algoritmo apresenta melhor desempenho?

Seleção de variáveis



- 55 *features* relevantes selecionadas por meio da técnica BORUTA
- 43 *features* necessárias identificadas por meio de RFE (*Recursive Feature Elimination)*

Seleção de variáveis

Dimensão	Variáveis relevantes (BORUTA)	Conjunto mínimo (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**

Seleção de Algoritmos

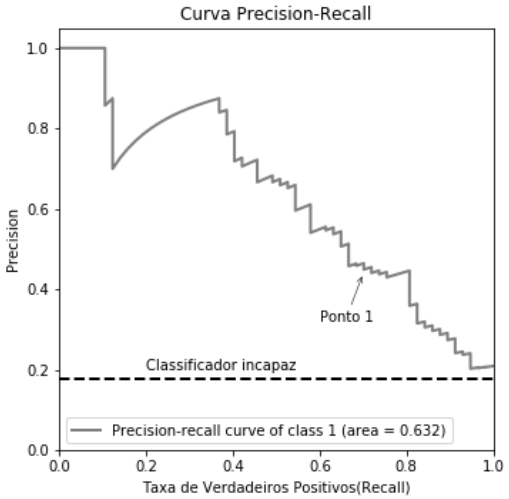
Algoritmos
Regressão Logística
Análise Discriminante Linear
K Vizinhos Próximos (KNN)
Naive Bayes
Máquina de Vetores de Suporte (SVM)
Perceptron Multicamada
XGBoost

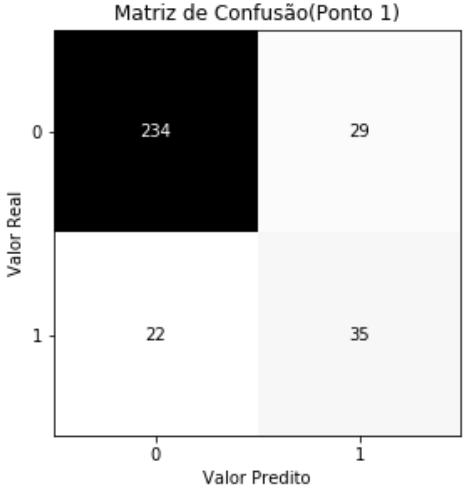
- Realizado um experimento, guiado por DSR, para avaliar a performance dos 7 algoritmos destacados, em suas configurações padrão
- Consideradas as 43 *features* mínimas
- Utilizada a base de Treinamento/Validação, com validação cruzada (k=10)
- Base de Teste utilizada para avaliar o desempenho

Resultado

Algoritmo	AUC PR	Recall	AUC ROC
Análise Discriminante Linear	**0,632**	0,614	**0,840**
Regressão Logística (RL)	0,611	**0,754**	0,823
Perceptron multicamada	0,570	0,491	0,815
Naive Bayes	0,529	0,421	0,775
Xgboost	0,528	0,649	0,761
Máquina de Vetores de Suporte (SVM)	0,495	0,526	0,771
K Vizinhos Próximos (KNN)	0,491	0,702	0,826
Classificador de Referência (RL)	0,372	0,404	0,683

Melhor algoritmo: LDA





O algoritmo LDA em conjunto com as *features* selecionadas mostrou desempenho muito superior ao algoritmo de referência, construído com RL e todas as *features* do *dataset.*

Algoritmo	AUC PR	Recall	AUC ROC
Análise Discriminante Linear	**0,632**	**0,614**	**0,840**
Regressão Logística (Referência)	0,372	0,404	0,683

DSR: Ciclo 2

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

Grid Search

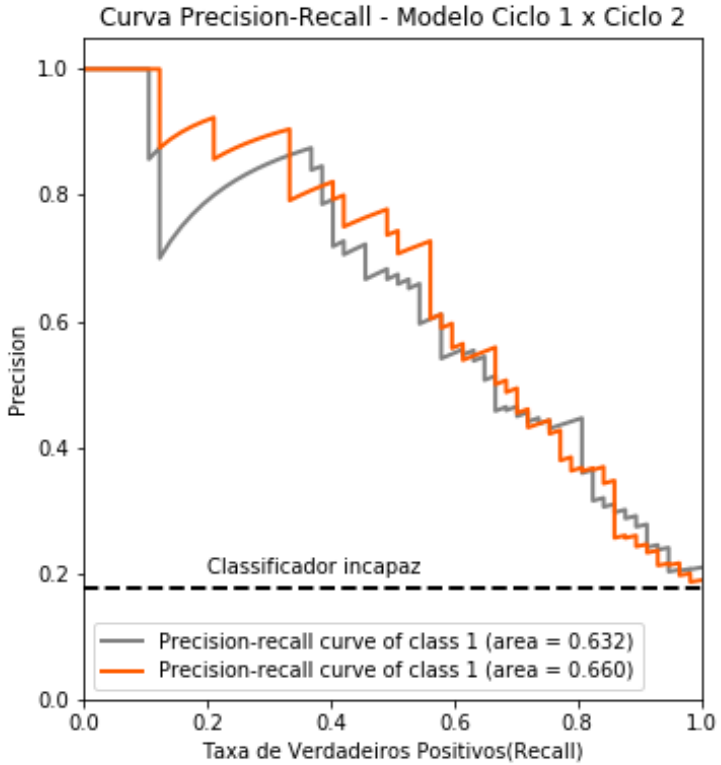
Algoritmo	Combinações de parâmetros avaliadas
Análise Discriminante Linear	351
Regressão Logística (RL)	50
Perceptron multicamada	864
Naive Bayes	9
Xgboost	9360
Máquina de Vetores de Suporte (SVM)	40
K Vizinhos Próximos (KNN)	72
Classificador de Referência (RL)	1

Realizou-se *grid search* em conjunto com validação cruzada para melhor ajuste de cada um dos algoritmos utilizados

Resultado

Algoritmo	AUC PR	Recall	AUC ROC
Máquina de Vetores de Suporte (SVM)	**0,660** 0,495	0,561 0,526	0,836 0,771
Regressão Logística (RL)	0,626 0,611	0,614 0,754	0,840 0,823
Análise Discriminante Linear	0,619 0,632	**0,842** 0,614	**0,843** 0,840
Xgboost	0,585 0,528	0,684 0,649	0,828 0,761
Perceptron multicamada	0,541 0,570	0,456 0,491	0,789 0,815
Naive Bayes	0,529 0,529	0,474 0,421	0,775 0,775
K Vizinhos Próximos (KNN)	0,436 0,491	0,649 0,702	0,807 0,826
Classificador de Referência (RL)	0,372	0,404	0,683

Resultado



Algoritmo	AUC PR	Recall	AUC ROC
Máquina de Vetores de Suporte (SVM)	**0,660**	0,561	0,836
Análise Discriminante Linear (Ciclo 1)	0,632	**0,614**	**0,840**
Regressão Logística (Referência)	0,372	0,333	0,683

- Parâmetros do SVM: C = 1.7 , kernel = rbf

Conclusão e contribuições

Conclusão: problema da pesquisa

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

- **modelo de AM supervisionada**
- **(SVM): C = 1.7** e **kernel = rbf**
- **43 *features*** selecionadas como minimamente necessárias
- **classe minoritária** (alto risco de óbito).
- ***AUC Precision-Recall* igual a 0,660** e *AUC-ROC* igual a 0,836.

Resumo Visual / Visual Abstract

Domínio Tecnológico: Classificar o risco de óbito de pacientes oncológicos de forma automática com os dados coletados na AGA por meio de Aprendizagem de Máquina

Entendimento do problema

Literatura referenciada aponta que existe capacidade preditiva de óbito na AGA

Instância do problema:

No IMIP, falta de suporte computacional dificulta priorização de centenas de pacientes oncológicos com base no risco individual.

Abordagem de validação

Dados divididos em treinamento/validação (80%) e teste (20%).

Solução:

Utilizando AM, em conjunto com os dados de 1600 pacientes já coletados (AGA) em um centro (IMIP), foi possível realizar classificação de forma automática.

Design da solução

Avaliamos e otimizamos 7 modelos de AM neste contexto. O de melhor desempenho foi o SVM.

Relevância: Para os centros oncológicos, ao desenvolvermos a base para uma ferramenta computacional capaz de classificar os pacientes. Para academia: ao demonstrarmos qual algoritmo e ajuste mais adequado a este contexto.

Rigor: Testados 7 algoritmos. De melhor desempenho foi o SVM, alcançando AUC PR igual à 0,660 e AUC ROC igual à 0,840.

Inovação: Utilização de AM com a métrica AUC PR, em conjunto com a AGA. Otimização do algoritmo. Detalhamento do pré-processamento .

Conforme descrito por (Engström et al., 2020)

Contribuições

1. **Modelo desenvolvido**
2. **Base de dados e Pré-processamento realizado**
3. **Identificação das variáveis relevantes e necessárias**
4. **Artigo aprovado**
   - CISTI ‘20, com o título *Machine Learning Applied to survival prediction of elderly cancer patients: Systematic Review*

Contribuições: variáveis necessárias

- O modelo apenas identifica os pacientes com alto risco de óbito, mas não diz o que deve ser feito pela equipe médica para modificar o desfecho.
- Dois grupos de variáveis identificadas:

- **Não Modificáveis**

Pertencem as dimensões **tipo de câncer**, **hábitos pregressos** e **estado mental**. Embora sejam utilizadas no modelo, não podem ser alteradas

- **Modificáveis**

Por exemplo: **avaliação nutricional**, **qualidade de vida**, **resultados clínicos** e **polifarmácia**. Essas variáveis podem ser alteradas pela equipe multidisciplinar de acompanhamento para reverter o desfecho

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

Limitações da pesquisa

- **O tamanho da base de dados** disponível para estudo (1600 pacientes)
- Não avaliou-se todos os tipos de algoritmos disponíveis, notadamente não dedicamos muito tempo **ao design de redes neurais**
- **Não utilizamos redes neurais profundas**
- **Não aplicamos a pesquisa em campo**

Trabalhos futuros

- **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 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)

Obrigado!

Perguntas?

Contatos

tblacerda@outlook.com

https://github.com/tblacerda/Mestrado_CESAR

Utilização de aprendizagem de máquina na classificação de risco de pacientes oncológicos

Tiago Beltrão Lacerda

Orientadora: Ana Paula Cavalcanti Furtado, D.Sc (CESAR.School / UFRPE)
Coorientadora: Dra. Jurema Telles de Oliveira Lima, D.Sc (IMIP)

Mestrado profissional em engenharia de software
Recife - 25/06/2020