UNIVERSIDADE ESTADUAL PAULISTA "JÚLIO DE MESQUITA FILHO"
FACULDADE DE CIÊNCIAS - CAMPUS BAURU
DEPARTAMENTO DE COMPUTAÇÃO
BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

CHRISTIAN LAURENCE ALMEIDA BARRY

MODELOS DE APRENDIZADO DE MÁQUINA BASEADOS EM KERNEL PARA ANÁLISE DE SOBREVIVÊNCIA

BAURU
Novembro/2025

CHRISTIAN LAURENCE ALMEIDA BARRY

MODELOS DE APRENDIZADO DE MÁQUINA BASEADOS EM KERNEL PARA ANÁLISE DE SOBREVIVÊNCIA

Trabalho de Conclusão de Curso do Curso de Bacharelado em Ciência da Computação da Universidade Estadual Paulista “Júlio de Mesquita Filho”, Faculdade de Ciências, Campus Bauru.
Orientador: Prof. Dr. João Paulo Papa
Coorientador: Danilo Samuel Jodas (PhD.)

BAURU
Novembro/2025

B279m

Barry, Christian Laurence Almeida

Modelos de aprendizado de máquina baseados em kernel para análise de sobrevivência / Christian Laurence Almeida Barry. -- Bauru, 2025

61 p. : il., tabs.

Trabalho de conclusão de curso (Bacharelado - Ciência da Computação) - Universidade Estadual Paulista (UNESP), Faculdade de Ciências, Bauru

Orientador: João Paulo Papa

Coorientador: Danilo Samuel Jodas

1. Ciência da computação. 2. Aprendizado de máquinas. 3. Análise do tempo de falha. I. Título.

Sistema de geração automática de fichas catalográficas da Unesp. Dados fornecidos pelo autor(a).

Christian Laurence Almeida Barry

Modelos de Aprendizado de Máquina Baseados em Kernel para Análise de Sobrevivência

Trabalho de Conclusão de Curso do Curso de Bacharelado em Ciência da Computação da Universidade Estadual Paulista “Júlio de Mesquita Filho”, Faculdade de Ciências, Campus Bauru.

Banca Examinadora

Danilo Samuel Jodas (PhD.)
Coorientador
Departamento de Computação
Faculdade de Ciências
Universidade Estadual Paulista "Júlio de Mesquita Filho"

Profa. Dra. Simone das Graças Domingues Prado
Departamento de Computação
Faculdade de Ciências
Universidade Estadual Paulista "Júlio de Mesquita Filho"

Prof. Dr. Kelton Augusto Pontara da Costa
Departamento de Computação
Faculdade de Ciências
Universidade Estadual Paulista "Júlio de Mesquita Filho"

Bauru, 10 de novembro de 2025.

Dedico este trabalho à minha família, que torceu por mim agora e sempre, e aos meus amigos, que tornaram meu tempo aqui tão especial.

Agradecimentos

Agradeço, primeiramente, à minha família: minha mãe, que sempre fez questão de acompanhar cada momento do meu dia-a-dia, me apoiando nos momentos difíceis e me celebrando nas conquistas, minha maior confidente quando as coisas ficavam pesadas; meu pai, que me guiou e me guia naquilo que ainda não sabia e naquilo que ainda não sei e em quem confio para informar minhas decisões; minha irmã, que sempre me faz consciente do quanto minha falta é sentida em casa, e que me motiva a ser alguém digno de admiração; meus avós, os membros da minha torcida particular que sempre me acompanharam de perto, mesmo de longe; meu tio Bruno, cuja amizade valorizo muito e que sempre está disposto a oferecer um sorriso e leveza.

Agradeço aos meus amigos: Fábio, Franco, Rufino, Vinícius, Sofia, Daniel, Leonardo, Yoshio, Luana, Luciano, Henrique e Eduardo, os membros da família que encontrei em Bauru. Me deram confiança quando me faltou, me levaram a evoluir exponencialmente enquanto pessoa e iluminaram os momentos bons e difíceis que compartilhamos em nossa jornada coletiva. Assim, se tornaram o motivo principal pelo qual posso considerar os anos que passamos juntos os melhores da minha vida.

Agradeço aos integrantes da PROTIVA: meus companheiros de equipe em meus times, meus coaches Luis Morelli e Roberta Spolon e Nicolas e Arissa, pelas aulas. Igualmente, agradeço a todos os integrantes do Carrossel Caipira. Ambos juntos me possibilitaram experiências inesquecíveis e acrescentaram peças fundamentais à minha formação.

Agradeço a todos os professores que me ensinaram ao longo da graduação, que me transformaram em um profissional a partir de um completo novato. Agradeço em específico ao professor Wilson Yonezawa, que me convidou a fazer parte da PROTIVA, e aos professores Higor Amario de Souza e Roberta Spolon, que assumiram o projeto quando parecia que ele poderia deixar de existir. Da mesma forma, agradeço aos professores Renê Pergoraro, fundador do Carrossel Caipira, e Clayton Pereira, que, hoje em dia, gerencia o projeto.

Por fim, agradeço ao professor João Paulo Papa, que me abriu as portas para a pesquisa ao me dar a oportunidade de fazer uma iniciação científica. Agradeço a Marcos Cleison, Guilherme Brandão e Danilo Jodas por me ajudarem e me orientarem ao longo desse processo, sempre com paciência mesmo em momentos em que eu não a merecia. Agradeço, também, à Petrobras por oferecer a bolsa para a produção do projeto e os dados necessários para sua elaboração e a do presente trabalho.

“Don’t worry, dude. Someone out there wants you. Promise.”

— **Susie**, em *DELTARUNE*, **Cap. 3**, de **Toby Fox** (2025).

Resumo

A análise de sobrevivência é um ramo da estatística essencial para a análise de dados de tempo até o evento, porém, métodos tradicionais têm dificuldade com dados complexos e não-lineares. Nesse paradigma, os modelos de aprendizado de máquina surgiram como soluções consolidadas para melhorar a análise de sobrevivência. No entanto, tais métodos exigem cuidadosa otimização de hiperparâmetros, um processo computacionalmente intensivo que dificulta implementações práticas. Esse trabalho enfrenta essas limitações ao propor três novos estimadores baseados em kernels desenvolvidos a partir da integração do estimador de Beran com dois algoritmos não-paramétricos do aprendizado de máquina: uma variação não-paramétrica do algoritmo de k vizinhos mais próximos (kNN) e o k*NN-supervised Optimum-Path Forest* (OPF-kNN). Ambos os métodos aprendem automaticamente o número ótimo de vizinhos ao longo de suas execuções, no caso do OPF-kNN, até um número máximo definido manualmente, evitando a maior parte do ajuste paramétrico. Através de uma avaliação abrangente em três conjuntos de amostras, um de *Downhole Safety Valves* empregadas em poços de petróleo, outro de pacientes sendo tratados para insuficiência cardíaca e, o terceiro, de pacientes passando por tratamento para câncer de próstata, foi constatado que os modelos desenvolvidos se aproximam ou superam a capacidade preditiva de modelos tradicionais dependentes de hiperparâmetros, caracterizando ferramentas robustas para análise de sobrevivência que lidam efetivamente com distribuições complexas e não-lineares sem requerer calibração específica para cada conjunto de dados.

Palavras-chave: Estimador de Beran; Análise de Sobrevivência; Análise de Confiabilidade; Confiabilidade para Engenharia; Hiperparâmetros; Aprendizado de Máquina.

Abstract

Survival analysis is a branch of statistics essetial to time-to-event analysis, but traditional methods struggle with complex, non-linear data. In this environment, machine learning models emerged as consolidated solutions to improve survival analysis. However, these methods demand careful hyperparameter tuning, a computationally intensive process that serve as an obstacle to practical implementations. This work addresses these limitations by proposing three new kernel-based estimators developed through the integration of the Beran estimator with two parameterless machine learning algorithms: a parameterless k-Nearest Neighbors (kNN) variation and kNN-supervised Optimum-Path Forest (OPF-kNN). Both of these methods automatically learn the optimal number of neighbors, up to a manually defined maximum amount, in OPF-kNN's case, avoiding most manual hyperparameter tuning. Through comprehensive evaluation of samples from three datasets, one of Downhole Safety Valves employed in oil wells, another of patients being treated for heart disease and a third of patients being treated for prostate cancer, it was asserted that our parameter-less approach maintains or surpasses the predictive capability of survival functions over traditional hyperparameter-dependent methods, providing a robust and adaptable tool for survival analysis that effectively handles complex, nonlinear data distributions without requiring dataset-specific calibration.

Palavras-chave: Beran Estimator; Survival Analysis; Reliabilty Analysis; Reliabilty for Engineering; Hyperparameters; Machine Learning.

Lista de figuras

Lista de tabelas

Sumário

1 Introdução

A análise de sobrevivência é uma ramo da estatística que visa prever o tempo até a ocorrência de algum evento crítico, utilizado nas áreas da engenharia (DIAMOUTENE et al., 2016), para prever falhas de equipamentos, e da medicina (LIMONGI et al., 2024), para prever o óbito ou recidiva de doenças em pacientes. A análise de sobrevivência tem sido abordada em diversos trabalhos na literatura, tanto do ponto de vista de estatística clássica (COLOSIMO; GIOLO, 2024) quanto da utilização de aprendizado de máquina (BENDER et al., 2020). Uma característica essencial da análise de sobrevivência é o tratamento da presença de dados censurados, uma característica única e essencial desta metodologia. A censura ocorre quando um evento de interesse não é observado para todos os indivíduos durante o período de estudo. Por exemplo, um paciente pode abandonar o acompanhamento médico, um equipamento pode ainda estar funcionando ao final da observação, ou um participante pode experimentar um evento diferente do que está sendo estudado. Em vez de descartar essas informações incompletas de incidência de eventos, a análise de sobrevivência as incorporá-las de forma inteligente, reconhecendo que o evento não ocorreu até determinado momento, o que ainda fornece informação valiosa. A capacidade de trabalhar adequadamente com observações censuradas permite extrair informações relevantes que seriam perdidas em análises estatísticas convencionais, maximizando o aproveitamento dos dados disponíveis e proporcionando estimativas mais precisas e confiáveis.

Na engenharia de confiabilidade, a análise de sobrevivência desempenha papel fundamental ao permitir a modelagem do tempo de vida de componentes, sistemas e equipamentos, o que permite fornecer subsídios para decisões estratégicas sobre manutenção, garantia e substituição. Através de funções como a taxa de falha, função de sobrevivência e tempo médio até a falha, é possível identificar padrões de degradação, prever comportamentos futuros e otimizar políticas de manutenção preventiva e preditiva. Esta abordagem é essencial para indústrias onde a confiabilidade é crítica, tais como no setor aeroespacial, nuclear, automotivo e petroquímica.

A análise de sobrevivência depende de dois pilares: a função de sobrevivência e a função de risco, que indicam a probabilidade de sobrevivência para além de um determinado tempo e o risco instantâneo de ocorrência de um evento, respectivamente. Métodos como o modelo de Cox (HARRELL JR; HARRELL, 2015), Weibull (MURTHY; XIE; JIANG, 2004), Exponencial (MARSHALL; OLKIN, 1967) e *Log-normal* (CROW; SHIMIZU, 1987) são ferramentas robustas para esse tipo de análise, mas que dependem de presunções específicas para estimar funções de sobrevivência. Sendo assim, surgiram, no âmbito do aprendizado de máquina, técnicas como a *Random*

Survival Forest (ISHWARAN et al., 2008) para trabalhar com relações não-lineares e sem parâmetros estritos, melhorando as previsões para uma ampla gama de situações (WANG; LI; REDDY, 2019). No contexto de análise de sobrevivência, covariáveis são características adicionais que auxiliam na dinâmica da função de sobrevivência em algum momento específico do tempo, permitindo modelar efeitos específicos que podem ter impacto na probabilidade de sobrevivência ao longo do tempo.

Para dados que não seguem uma distribuição específica, empregam-se métodos não-paramétricos, com destaque ao Estimador de Beran (BERAN, 1981) e Kaplan-Meier (KAPLAN; MEIER, 1958). Esses estimadores se sobressaem por serem não-paramétricos e, no caso do Beran, em suas capacidades de acomodar os efeitos das covariáveis para esses conjuntos de dados, melhorando as avaliações das relações entre as características das amostras e os tempos de sobrevivência.

O estimador de Beran, por exemplo, opera baseado em *kernels* que modelam a relação entre as amostras do conjunto. Entretanto, esses *kernels* têm seu funcionamento prejudicado quando aplicados a dados com altas dimensões ou com relações complexas e não-lineares entre suas covariáveis. Frente a tal obstáculo, uma variação do estimador foi proposta: o Estimador de Beran com *kernels* Neurais (BENK) (KIRPICHENKO et al., 2024), isto é, um Estimador de Beran baseado em *kernels*, em que cada *kernel* é formado por uma rede neural. Contudo, além da necessidade de uma extensiva otimização dos hiperparâmetros, tanto para a quantidade de subredes neurais que formam os *kernels* quanto para os subconjuntos de dados usados durante o treinamento do modelo, o custo computacional de implementar as diversas redes neurais se mostra bastante elevado comparado aos modelos tradicionais.

Frente a tal paradigma, mostra-se interessante que hajam modelos de aprendizado de máquina não-paramétricos que dispensem a necessidade de ter seus hiperparâmetros otimizados, mas cujo desempenho seja comparável, ou melhor, ao dos modelos para os quais isso é necessário. Belle et al. (2011) fazem uma comparação entre abordagens baseadas em ranqueamento e regressão para análise de sobrevivência e demonstram que os modelos que incorporam regressão, que são baseados em *kernels*, tem desempenho consideravelmente melhor do que os que incorporam ranqueamento. A partir disso, pode-se concluir que, mesmo que os *kernels* tradicionais tenham dificuldade quando os dados são não-lineares e de alta dimensionalidade, métodos que os incorporam ainda mostram desempenho superior ao de seus equivalentes em tais situações.

Sendo assim, chama atenção um novo modelo baseado em *kernels* que minimize a otimização de hiperparâmetros. O algoritmo de k-Vizinhos mais Próximos (k-NN), símples porém robusto, tem seu desempenho afetado por mudanças ao número de vizinhos, o parâmetro k. Para resolver esse problema, surgem os modelos *Para-*

meterless k-Nearest Neighbors (PL-kNN) e o *k-NN supervised Optimum-Path Forest* (OPF-kNN) (PAPA; FALCAO; SUZUKI, 2009), ambos dos quais aprendem o valor ótimo de k ao longo de sua execução, até, no caso do OPF-kNN, um máximo pré-definido. Com isso, os dois modelos eliminam ou minimizam a necessidade da otimização de hiperparâmetros, se tornando bons candidatos para novas funções *kernel* aplicáveis ao estimador de Beran.

Portanto, este trabalho propõe o desenvolvimento de modelos de aprendizado de máquina para análise de sobrevivência que sejam baseados em *kernel* e que tenham poucos, idealmente nenhum, hiperparâmetros a serem otimizados, poupando tempo e recursos que essa etapa precursora demanda na implementação dos métodos já existentes.

1.1 Contextualização do problema de pesquisa

Modelos de aprendizado de máquina para análise de sobrevivência, tais como *Random Survival Forest* (RSF), *Survival Support Vector Machines* (SSVM) e Estimador de Beran com *Kernels* Neurais (BENK), apesar de robustos, apresentam algumas complicações no que diz respeito à sua parametrização, visto que tais modelos requerem uma etapa prévia de otimização de hiperparâmetros. No caso do RSF, alguns hiperparâmetros incluem a quantidade de árvores, a quantidade de variáveis por nó das árvores e um número mínimo de amostras por nó folha e, para o BENK, devem ser otimizadas tanto a quantidade de subredes neurais que formam os *kernels* quanto os subconjuntos de dados usados durante o treinamento do modelo. Por sua vez, é necessária para o SSVM a otimização dos fatores de regularização e hiperparâmetros associados aos *kernels* da SVM.

Esses hiperparâmetros impactam o desempenho, capacidade de generalização e eficiência do modelo e são fundamentais para a capacidade preditiva. Porém, o processo manual de sua otimização por tentativa e erro, como mencionado por Bischl et al. (2023), é um processo demorado e irreproduzível, além de computacionalmente intensivo.

A execução dos algoritmos em si, por sua vez, são, também, custosos. Huang et al. (2023), em sua revisão de métodos de aprendizado de máquina para análise de sobrevivência, concluiram que o RSF pode ser lento para conjuntos de dados extensos, ao passo que o treinamento de numerosas redes neurais que ocorre no modelo BENK requer recursos computacionais consideráveis.

Tendo isso em vista, a necessidade da otimizar hiperparâmetros faz com que os algoritmos de aprendizado de máquina, que muitas vezes já são custosos, fiquem ainda mais computacionalmente intensivos, prejudicando um tempo de execução já

extenso. No entanto, modelos que não apresentam essa necessidade são raros no paradigma da análise de sobrevivência.

1.2 Justificativa

Para reduzir a carga computacional e consumo de tempo da tarefa de otimização de hiperpârametros, diversos métodos foram concebidos, como a otimização bayesiana (SNOEK; LAROCHELLE; ADAMS, 2012) e *bandit-based configuration evaluation* (LI et al., 2022). No entanto, de acordo com Falkner, Klein e Hutter (2018), devido aos longos tempos de treinamento de modelos estado-da-arte, a otimização bayesiana é tipicamente computacionalmente inviável, enquanto as abordagens *bandit-based* não convergem rapidamente às melhores configurações. Logo, um modelo sem hiperparâmetros é um que economiza tempo e recursos investidos na otimização automática, ao mesmo tempo que remove qualquer enviesamento e erro humano, algo que frequentemente ocorre num processo de otimização manual do modelo.

Isso é especialmente relevante quando se trata de modelos com diversos hiperparâmetros a serem otimizados. Vários dos algoritmos mais popularmente aplicados à análise de sobrevivência, como SSVM, RSF, e BENK, compartilham dessa característica, tornando a área uma das mais afetadas pelo problema vigente e, portanto, uma das mais beneficiadas pela sua solução.

A maior agilidade obtida é de impacto ainda maior quando se considera as áreas de aplicação da análise de sobrevivência. No âmbito da medicina, prever o tempo até um resultado de interesse é de importância crítica para pesquisas clínicas (LEE; GO, 1997), enquanto a previsão da falha de um equipamento, na engenharia, traz uma ampla variedade de benefícios, como manutenção preditiva e substituições que antecipam o problema. (PAPATHANASIOU; DEMERTZIS; TZIRITAS, 2023).

Contudo, uma economia de tempo é relevante apenas se o desempenho do modelo for competitivo. Uma semelhança de resultados entre um modelo que requer otimização de parâmetros e um que a descarta já foi apresentada por Jodas et al. (2022) ao propor o modelo *Parameterless* k-*Nearest Neighbors* (PL-kNN), que contorna essa etapa ao determinar a quantidade ótima de vizinhos com base na distribuição de dados. Além disso, destaca-se o modelo baseado em grafos *Optimum-Path Forest* (OPF) desenvolvido por Papa, Falcao e Suzuki (2009), cujas funcionalidades contemplam tarefas de classificação supervisionada que não necessita de otimização de hiperparâmetros. Tal evidência mostra que os modelos sem otimização podem se comparar ou até superar seus competidores baseados em hiperparâmetros.

Visto isso, este trabalho consiste em reduzir a carga computacional dos modelos de aprendizado de máquina de análise de sobrevivência aplicada à falha de

equipamentos e a eventos médicos. Isso agilizaria pesquisas médicas e preveria falhas de equipamentos com mais prontidão, ajudando a melhorar a qualidade de vida e a previnir acidentes.

1.3 Objetivos

1.3.1 Objetivo Geral

Propor novos modelos de aprendizado de máquina baseados em *kernel* voltados à análise de sobrevivência que sejam não-paramétricos e que não necessitem de otimização de hiperparâmetros, para remover essa etapa do desenvolvimento desses modelos.

1.3.2 Objetivos Específicos

- Avaliar os modelos tradicionais de análise de sobrevivência, baseados em estatística tradicional.
- Explorar os modelos de análise de sobrevivência baseados em aprendizado de máquina, particularmente no que tange ao uso de covariáveis.
- Explorar modelos baseados em *kernels* para a análise de sobrevivência.
- Estudar técnicas e algoritmos aplicáveis ao problema.
- Desenvolver novos modelos de *kernel* não-paramétricos para utilização em aprendizado de máquina visando a análise de sobrevivência.
- Comparar o desempenho dos novos modelos com os principais modelos de análise de sobrevivência, tanto os baseados em estatística tradicional, quanto os baseados em aprendizado de máquina.

2 Referencial Teórico

Essa seção apresenta os principais fundamentos relacionados à análise de sobrevivência e sua importância no contexto da medicina e engenharia. Serão apresentados os conceitos de censura de dados e a formulação matemática dos principais modelos paramétricos e não-paramétricos aplicados à análise de sobrevivência e análise de confiabilidade. Os conceitos abordados representam os pilares fundamentais para o entendimento da modelagem de métodos avançados que transcendem a abordagem estática tradicional, particularmente no que tange ao uso de aprendizado de máquina para análise de sobrevivência.

2.1 Análise de Sobrevivência

A análise de sobrevivência se trata de um ramo de estatística que busca determinar o de tempo até a ocorrência de um evento crítico ou de interesse (BENDER et al., 2020). Esse estudo tem aplicações principalmente nos campos da medicina, cujo objetivo é determinar o tempo até a recidiva ou melhora dos pacientes que fazem parte do estudo, e da engenharia, onde estuda o tempo até a falha de equipamentos. Nesse segundo caso, ela recebe o nome de análise de confiabilidade (COLOMBO et al., 2022; LOUZADA et al., 2020).

No entanto, existem diversos outros contextos em que é possível aplicar a análise de sobrevivência, como, por exemplo, a estimativa do tempo até o encerramento de um programa de fidelidade ou alguma assinatura (ARAÚJO, 2022). Outros exemplos de possíveis aplicações podem ser encontrados na tabela 1. De modo geral, pode-se observar que em qualquer âmbito em que hajam eventos que tem um início e um fim, ou que haja mudanças de estados ao longo do tempo, pode ser aplicada a análise de sobrevivência.

Tabela 1 – Exemplos de aplicações da análise de sobrevivência

Campo de aplicação	**Início**	**Fim**	**Censura**
Confiabilidade (COLOMBO et al., 2020; LOUZADA et al., 2019)	Instalação do equipamento, início da operação, retorno de uma manutenção	Falha do equipamento	Não-falha ou substituição não-relacionada a falha do equipamento
Finanças (CZADO; RUDOLPH, 2002; FARD et al., 2016)	Contratação de um empréstimo, seguro	Quitação do empréstimo, acionamento do seguro	Encerramento da dívida por outros motivos, óbito.
Ciências sociais (AMERI et al., 2016)	Início de um emprego	Mudança de função	Demissão
Medicina (ZHU; YAO; HUANG, 2016; KIM et al., 2019)	Início do acompanhamento de uma doença	Óbito	Óbito por causas alheias à doença acompanhada, perda de acompanhamento

Fonte: Elaborado pelo autor

Um dos maiores obstáculos para os estudos de análise de sobrevivência provém da censura, isto é, uma informação incompleta que surge quando o evento de interesse não ocorre ao longo do tempo de estudo (COLOSIMO; GIOLO, 2024). Alguns exemplos incluem, na confiabilidade, um equipamento que foi trocado ou não falhou e, na medicina, um paciente que não apresentou recidiva durante o acompanhamento ou foi a óbito por causas alheias ao objeto de estudo (COLOMBO, 2023; ZHU; YAO; HUANG, 2016).

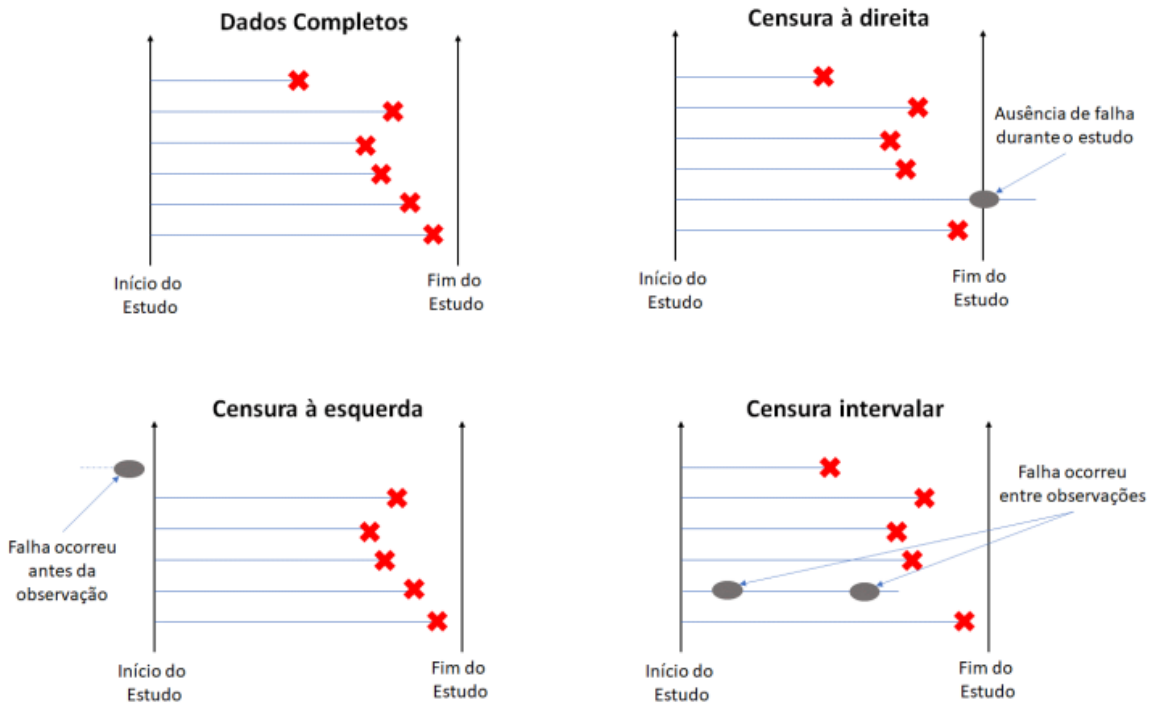
Apesar de incompleta, como estabelece um tempo mínimo em que a amostra sobreviveu durante o período de observação, a amostra censurada não pode ser descartada sem comprometer a qualidade do estudo, algo que torna o tratamento da censura um dos desafios fundamentais para o subdomínio estatístico (SPOONER et al., 2020). Nesse contexto, surgem três tipos de censura: a censura à esquerda, intervalar, e à direita, descritas na tabela 2 e representadas visualmente na figura 1. Segundo (LAWLESS, 2011; LEE; WANG, 2003; MEEKER; ESCOBAR; PASCUAL, 2021; MODARRES; KAMINSKIY; KRIVTSOV, 2016):

Tabela 2 – Explicação dos diferentes tipos de censura.

Tipo de censura	**Descrição**
À esquerda	Ocorrência do evento antes do início do estudo
Intervalar	Existência de um intervalo no período de estudo em que não se tem informações acerca do evento
Tipo I à direita	Não-ocorrência do evento durante o período de acompanhamento
Tipo II à direita	Condução do estudo até que um determinado número menor que o total de amostras de falhas ocorra

Fonte: Elaborada pelo autor.

Figura 1 – Tipos de dados censurados



Fonte: Colombo (2023)

2.1.1 Conjunto de dados

Formalmente, uma amostra ou observação em análise de sobrevivência é dada por uma tupla $\{ t_i, \delta_i, \overrightarrow{x_i} \}$, onde t_i é o tempo de evento observado para a observação i, δ_i é o indicador de censura que assume valor 0 quando houve censura ou 1 em caso contrário e $\vec{x}_i$ é o vetor m-dimensional de covariáveis $\{x_i^1, x_i^2, \ldots, x_i^m\}^T$ em que cada elemento deste vetor é uma característica da observação (BENDER et al., 2020).

Dessa forma, o conjunto de dados é constituído por n observações referentes ao tempo até a ocorrência de um evento de interesse em indivíduos, sistemas ou

unidades sob estudo. Trata-se de um conjunto de dados heterogêneo, pois nem todos os eventos ocorreram sob as mesmas condições ou com as mesmas características. Isso significa que, dentro do conjunto, podem existir diferenças relacionadas a fatores como condições ambientais, características demográficas, tratamentos recebidos, contextos de uso ou exposição, entre diversas outras variáveis que podem influenciar o tempo até o evento.

2.1.2 Funções matemáticas da análise de sobrevivência

Na análise de sobrevivência, busca-se estudar o tempo até a ocorrência de um evento de interesse, como óbito, recaída de uma doença ou falha de um sistema, considerando um conjunto de condições e fatores que podem influenciar este tempo. O objetivo é quantificar a probabilidade de sobrevivência de um indivíduo ou unidade alvo durante um período específico (LAWLESS, 2011).

A função que nos traz essa informação é a função de sobrevivência, $S(t)$, que retorna a probabilidade do evento de interesse não ter ocorrido até o tempo t, ou seja, formalmente, indica a probabilidade até o evento de interesse no tempo T ser maior do que no tempo t, dado que $T \leq t$.

Assim, se $f(t)$ é a função densidade de probabilidade do tempo até o evento, a função de sobrevivência pode ser escrita como (Eq. 1):

$$S(t) \;=\; \mathbb{P}(T \geq t) \;=\; \int_t^{\infty} f(s)\, ds \tag{1}$$

A função de sobrevivência possui as seguintes propriedades:

- $S(t) \geq 0$ para todo $t \geq 0$;
- $S(0) = 1$;
- $S(\infty) = \lim_{t\to\infty} S(t) = 0$;
- $S(t)$ é monotonicamente decrescente.

Além da função de sobrevivência, outras funções matemáticas também são usadas na análise. A primeira delas é a **função densidade de probabilidade** $f(t)$, que descreve a distribuição do tempo até o evento. Sua definição formal é dada por (Eq. 2):

$$f(t) \;=\; \lim_{\Delta t\to 0} \frac{\mathbb{P}(t \leq T < t + \Delta t)}{\Delta t} \tag{2}$$

Nessa mesma linha, pode-se citar a **função distribuição acumulada** do tempo até o evento, denotada por $F(t)$, que age como complemento da função de sobrevivên-

cia ao indicar a probabilidade que o evento tenha ocorrido até o tempo t, correspondendo, formalmente, à probabilidade de que o tempo T até o evento seja menor que t (Eq. 3):

$$F(t) = 1 - S(t) = \mathbb{P}(T < t) = \int_0^t f(s)\, ds \tag{3}$$

Além dessas, as demais funções comuns se tratam de funções condicionais, que se baseam na não-ocorrência do evento até um certo instante t, ou seja, mostra a probabilidade do evento ocorrer dado que ele não tenha acontecido até esse mesmo instante.

Uma destas funções é a **função risco instantâneo** (ou *hazard rate*), definida como a densidade de probabilidade condicional (Eq. 4):

$$h(t) = \lim_{\Delta t \to 0} \frac{\mathbb{P}(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t} = \frac{f(t)}{S(t)} \tag{4}$$

A partir dela temos a função risco acumulado (Eq. 5):

$$H(t) = \int_0^t h(s)\, ds \tag{5}$$

E, a partir dessa última, podemos obter diretamente a função de sobrevivência (Eq. 6):

$$S(t) = e^{-H(t)} \tag{6}$$

Por fim, temos a **sobrevivência condicional**, que representa a probabilidade de um indivíduo sobreviver por mais t unidades de tempo dado que já sobreviveu por s unidades. Assim, ela é dada por (Eq. 7):

$$S(t \mid s) = \frac{S(t+s)}{S(s)} \tag{7}$$

2.2 Principais métodos estatísticos clássicos

As funções apresentadas na última seção podem ser obtidas por meio de uma ampla gama de métodos estatísticos, divididos principalmente entre métodos paramétricos, que assumem uma certa distribuição estatística dos dados, essa distribuição variando de modelo para modelo, e modelos não-paramétricos, que não apresentam essa mesma característica.

O objetivo desses modelos é, a partir dos dados históricos de vida dos indivíduos estudados, estabelecer uma função estimada para $S(t)$ denominada de $\hat{S}(t)$.

Os modelos paramétricos, que objetivam descobrir o valor dos parâmetros, usam diferentes distribuições de probabilidade para ajustar a função $\hat{S}(t)$ como uma função contínua obtida a partir da regressão dos tempos de sobrevivência. A partir desse ajuste, se torna possível avaliar a probabilidade de sobrevivência para tempos fora do conjunto de dados utilizado para fazer o ajuste. É mais comum o uso destes na engenharia do que na medicina.

Enquanto isso, os modelos não-paramétricos, ao não assumirem alguma distribuição estatística previamente, evitam hipóteses a priori sobre este comportamento, já que essas, como podem não ser necessariamente válidas, podem levar a resultados errôneos sem devida justificativa (CHAKRABORTY; TSOKOS, 2021). Em geral, se baseam na função empírica (VAJARGAH, 2021), dada por (Eq. 8):

$$H_n(t) = \frac{i}{n} \tag{8}$$

onde i é o número de falhas ocorridas até o tempo t e n é o tamanho da amostra. Essa função indica a taxa de falha, que nos leva a função de sobrevivência facilmente por meio de $1 - H(t)$. A seguir estão descritos os principais modelos paramétricos e não-paramétricos:

2.2.1 Modelos paramétricos

2.2.1.1 Distribuição exponencial

Esse modelo se caracteriza por sua simplicidade, possuindo apenas um parâmetro. As equações (9) e (10) mostram, respectivamente, a expressão que define a distribuição exponencial e a função de sobrevivência obtida a partir dela, segundo (RASHEED, 2023):

$$f(t) = \lambda e^{-\lambda t} \tag{9}$$

$$S(t) = 1 - F(t) = e^{-\lambda t} \tag{10}$$

Aqui, $f(t)$ representa a distribuição exponencial, $S(t)$ indica a função de sobrevivência, $h(t)$, a taxa de falha e, por fim, $S(t \mid s)$ a função de sobrevivência condicional. Assim, podem ser observadas duas peculiaridades da exponencial: primeiramente, a taxa de falha se reduz a uma constante, fazendo com que ela não dependa do tempo, ao mesmo tempo que a função de sobrevivência condicional depende apenas do tempo total t e não do tempo s que o indivíduo já sobreviveu.

2.2.1.2 Distribuição Weibull

A distribuição Weibull se caracteriza pela maior flexibilidade em relação à exponencial, não sendo restrita a uma taxa de falha constante. Porém, essa taxa de falha ainda é restrita a um comportamento monotonicamente decrescente, não sendo capaz de representar taxas de falha que variam entre crescência e decrescência ao longo do tempo.

Seus dois parâmetros são o parâmetro de escala, representado por η e indicativo do valor de tempo onde 63,2% das falhas ocorrem, e o parâmetro de forma, representado por β e indicativo do comportamento da função (decrescente, para $\beta < 1$, constante, para $\beta = 1$ ou crescente, para $\beta > 1$). Seguem, nas equações (11) e (12), respectivamente, as funções de análise de sobrevivência obtidas a partir da distribuição de Weibull, segundo (LAI; MURTHY; XIE, 2006):

$$f(t) = \frac{\beta}{\eta}\left(\frac{t}{\eta}\right)^{\beta-1} e^{-\left(\frac{t}{\eta}\right)^{\beta}} \tag{11}$$

$$S(t) = 1 - F(t) = e^{-\left(\frac{t}{\eta}\right)^{\beta}} \tag{12}$$

2.2.1.3 Distribuição Log-normal

A distribuição log-normal apresenta certas características que a encaixam bem no contexto da análise de sobrevivência. Para começar, ela tem em seu domínio apenas em valores positivos de t, algo adequado quando se considera a inexistência de tempos de falhas negativos. Ademais, se mostra uma boa escolha em cenários onde os eventos de falha são causados por uma interação de diversos fatores que influenciam o tempo até o evento de interesse já que surge naturalmente como produto de diversas variáveis independentes.

A distribuição log-normal é caracterizada por dois parâmetros principais: μ, que representa a média do logaritmo dos tempos de falha, e σ, que indica o desvio padrão do logaritmo dos tempos de falha. As Equações (13) e (14) apresentam as principais funções matemáticas utilizadas na análise de sobrevivência para essa distribuição (GUPTA; KANNAN; RAYCHAUDHURI, 1997).

$$f(t) = \frac{1}{t\,\sigma\sqrt{2\pi}}\,\exp\left\{-\frac{1}{2\sigma^2}\big(\ln t - \mu\big)^2\right\}, \qquad t > 0,\ \sigma > 0, \tag{13}$$

$$S(t) = 1 - F(t) = \Phi\left(\frac{\mu - \ln t}{\sigma}\right). \tag{14}$$

2.2.2 Modelos não-paramétricos

2.2.2.1 Kaplan-Meier

O modelo não-paramétrico mais conhecido se trata do estimador de Kaplan-Meier, sendo esse também o mais utilizado em análise de sobrevivência com dados censurados (KAPLAN; MEIER, 1958). Se trata de uma função degrau baseada na função empírica, que desce de degrau nos tempos dos eventos de interesse (SCHOBER; VETTER, 2021).

O cálculo da sua função de sobrevivência exige que se ordene os tempos de evento T. Então, para cada tempo de interesse $t_j, 1 \leq j \leq k$, em que k é o tempo máximo avaliado, são determinados a quantidade de eventos que ocorreram em t_j, bem como os indivíduos ainda sob risco em t_j, isto é, os indivíduos para os quais o evento de interesse ainda não ocorreu e que não foram censurados até o tempo t_j. Essas duas quantidades são dadas por d_j e n_j, respectivamente.

Assim, a equação de Kaplan-Meier para a função de sobrevivência é dada pela equação (15):

$$\hat{S}(t) = \prod_{j:t_j<t} \left(\frac{n_j - d_j}{n_j} \right) = \prod_{j:t_j<t} \left(1 - \frac{d_j}{n_j} \right) \tag{15}$$

O funcionamento do estimador leva a um domínio divido em partes de quantidade igual à quantidade de eventos ocorridos. A ideia é que, para que um indivíduo sobreviva até o tempo t, ele deve ter sobrevivido a todos os intervalos anteriores.

É comum encontrar estudos utilizando o estimador de Kaplan-Meier como referência para o desempenho de novos modelos desenvolvidos (KIESSLING et al., 2023; SHOURABIZADEH et al., 2025; THORSEN-MEYER et al., 2022). Nesse estudo, os experimentos realizados seguirão esse padrão, observando a estimativa do Kaplan-Meier e seu intervalo de confiança para avaliar a performance dos novos modelos obtidos.

2.2.2.2 Estimador de Beran

O estimador de Beran (BERAN, 1981) é um método não-paramétrico focado na estimativa da função de sobrevivência condicional. Se trata de uma extensão do Kaplan-Meier que incorpora os efeitos das covariáveis em sua estimativa por meio da suavização por *kernel*. Isso o confere robustês e flexibilidade em relação aos demais métodos estatísticos, que não consideram as características individuais e, logo, não capturam os efeitos das covariáveis, assumindo que estes serão fixos.

A maior vantagem desse estimador está em sua flexibilidade e adaptabilidade

a diferentes conjuntos de dados, pois, além de acomodar os efeitos das covariáveis, seu status como método não-paramétrico também significa que ele não assume um distribuição específica dos dados, ao mesmo tempo que acomoda censura a direita. Essas características o tornam uma importante ferramenta quando existem relações complexas entre os dados, particularmente no que diz respeito às covariáveis.

O estimador de Beran para uma função de sobrevivência condicional $S(t \mid \mathbf{x})$ no tempo t considerando o vetor de covariáveis $\mathbf{x}$ é dado por (Eq. 16):

$$S(t \mid \mathbf{x}) = \prod_{T_i \leq t} \left(1 - \frac{K(x_i - \mathbf{x})\, \delta_i}{\sum_{j:T_j \geq T_i} K(x_j - \mathbf{x})} \right), \tag{16}$$

onde t é o tempo máximo observado, T_i representa o i-ésimo tempo observado, δ_i é o indicador de evento da i-ésima observação (1 se o evento ocorreu ou 0 se censurado), $\mathbf{x}_i$ denota o vetor de covariáveis para o i-ésimo indivíduo, e $K(\cdot)$ é uma função *kernel*. A condição $T_j \geq T_i$ representa todos os tempos remanescentes após T_i.

A performance do estimador é fortemente ligada a função de kernel, sendo escolhas comuns o kernel Guassiano e de Epanechnikov (LÓPEZ-CHEDA et al., 2017). No caso de uma função de densidade de kernel com largura de banda h, a banda é selecionada por métodos de *cross-validation*, porém, essa abordagem pode aumentar a complexidade computacional do algoritmo (LÓPEZ-CHEDA et al., 2017). Ademais, o uso de suavização por kernel em geral pode introduzir enviesamento de fronteira perto das extremidades da distribuição das covariáveis. Apesar disso, a aplicabilidade do estimador vem aumentando em análise de sobrevivência de larga escala.

2.3 Aprendizado de máquina

O aprendizado de máquina, do inglês *Machine Learning*, é um campo de estudo da computação que evoluiu do reconhecimento de padrões, fazendo parte da esfera da inteligência artificial. Em pesquisas recentes, os modelos de aprendizado de máquina se tornaram atraentes no âmbito da análise de sobrevivência por serem capazes de tratarem não-linearidades para fazer predições generalizadas sobre eventos futuros, mesmo a partir de dados históricos complexos (SANTOS, 2018).

Os modelos de aprendizado de máquina podem, em sua maioria, ser divididos em duas categorias: os que resolvem problemas de regressão, e os que resolvem problemas de classificação. Enquanto os problemas de regressão focam em ajustar uma curva para que modele o conjunto de dados através do ajuste dos seus pesos, retornando um valor quantitativo (HASTIE, 2009), os problemas de classificação buscam atribuir um rótulo a uma amostra entre determinadas categorias, chamadas classes e representadas por C_i. Nesse segundo tipo, o valor de retorno representa um rótulo

indicando a classe C_i à qual a observação sob análise pertence, ou intervalar, variando de $p_i \in [0, 1]$ com base na probabilidade da amostra fazer parte de uma das classes C_i (ALPAYDIN, 2020).

Há ainda a divisão entre modelos supervisionados e não-supervisionados (ALLOGHANI et al., 2020; KUHN; JOHNSON et al., 2013). Em suma, modelos supervisionados recebem dados rotulados durante a etapa de aprendizagem, relacionando suas covariáveis com o rótulo recebido, enquanto não-supervisionados recebem apenas as covariáveis e fazem a classificação a partir de um agrupamento com base na semelhança entre diferentes conjuntos de amostras, chamado *clustering* (HASTIE, 2009; JAMES et al., 2013).

A motivação por trás do emprego desses métodos é a melhora de qualidade das predições pela sua capacidade de modelar relações complexas entre os dados, apresentando uma evolução dos métodos não-paramétricos (ALSINA; CABRI; REGATTIERI, 2016). No entanto, a presença de censura nos conjuntos de dados referentes à análise de sobrevivência representa um grande desafio por se assemelharem a uma amostra não rotulada num problema de classificação, ou uma resposta desconhecida, em um problema de regressão (FARD et al., 2016).

É possível pensar em um problema de análise de sobrevivência como um problema de classificação, em que as classes são a ocorrência ou não de um evento de interesse, em um dado tempo t para uma determinada amostra (ZHONG; TIBSHIRANI, 2019). No entanto, como a classificação ocorre em um instante do tempo, não se pode fazer uso dessa abordagem para modelar a função de sobrevivência do indivíduo ao longo do tempo. Para isso, é necessário um modelo de regressão que entenda a sua respota Y como a probabilidade de sobrevivência baseada em um vetor X formado pelo tempo e pelas covariáveis do indivíduo (JAMES et al., 2013). Formalmente (Eq. 17):

$$Y \approx f(x) = f(t, \vec{x}) \tag{17}$$

em que $\vec{x}$ é o vetor de covariáveis e t, o tempo.

A metodologia para a implementação de um modelo de aprendizado de máquina segue os seguintes passos (LEE, 2019):

- Pré-processamento
- Separação aleatória do conjunto de dados em treinamento e validação
- Seleção de modelo de aprendizado
- Predição da resposta de interesse em dados de teste

- Avaliação dos modelos selecionados

O conjunto de treinamento se trata de um conjunto de amostras usadas na aprendizagem do modelo, enquanto o conjunto de validação é usado para avaliar o desempenho do algoritmo com base nas métricas adequadas (IZBICKI; SANTOS, 2020).

A seguir, estão descritos os modelos de aprendizado de máquina frequentemente usados em análise de sobrevivência de interesse para esse trabalho.

2.3.1 *Support Vector Machines*

O método *Support Vector Machines* (SVM) foi desenvolvido por Cortes & Vapnik (CORTES; VAPNIK, 1995) e é utilizado para classificação, criando uma superfície de separação entre classes através de um hiperplano que maximiza a distância entre as observações. Ele é eficiente em lidar com relações não lineares, utilizando funções *kernel*, além de ser útil para problemas de alta dimensionalidade (KHAN; ZUBEK, 2008).

O SVM também pode ser adaptado para problemas de regressão, sendo chamado de SVR (*Support Vector Regression*). Neste caso, o modelo é representado por uma função linear, onde a função kernel transforma o espaço das características em uma dimensão mais alta. A função utilizada para ajuste é uma função penalizada que busca minimizar a distância entre os pontos e a reta, controlando a margem de erro.

A equação (18) mostra a representação matemática do modelo de SVM:

$$f(x) = W \cdot \varphi(x) + b \tag{18}$$

onde $\varphi(x)$ é uma função kernel que transforma o espaço de características das entradas e um novo espaço de alta dimensão e W e b são parâmetros obtidos através do treinamento do modelo, minimizando a função custo exibida na equação (19):

$$\min_{w,b} J(w,b) \;=\; \frac{1}{2}\,\|w\|^2 \;+\; C\sum_{i=1}^{n}\big(\xi_i + \xi_i^*\big), \tag{19}$$

Essa função de custo pode ser interpretada da seguinte forma: o algoritmo busca o hiperplano que melhor ajusta os dados no espaço de características. A partir desse hiperplano, avalia-se a discrepância entre a predição e o alvo, $f(x) - y$. Amostras cuja discrepância fica abaixo de um erro aceitável, esse representado por ε, são consideradas bem ajustadas; as demais são mal ajustadas.

Como, em geral, não é possível ajustar perfeitamente todos os pontos, introduzem-se as *variáveis de folga* ξ_i, ξ_i^*, caracterizando a formulação de *soft margin* do SVM. De

maneira simplificada, ξ_i e ξ_i^* contabilizam violações além da margem/tubo aceitável e impõem uma penalização no funcional de custo. O parâmetro $C > 0$ controla o peso dessa penalização, regulando o compromisso entre a complexidade do modelo, representada pelo termo $\frac{1}{2}\|w\|^2$, e o erro de ajuste nas amostras (KHAN; ZUBEK, 2008).

2.3.2 *Random Survival Forests*

Random Forests (árvores aleatórias) são um método de aprendizado supervisionado que avalia novas amostras a partir de um conjunto de árvores de decisão formadas, por sua vez, com base no conjunto de treinamento (BREIMAN, 2001). Árvores de decisão são árvores construídas começando no nó raiz e subdividindo os nós com o objetivo de dividir o conjunto de dados em subconjuntos mais homogêneos (JAMES et al., 2013). Apesar da fácil implementação, muitas vezes apresentam desempenho abaixo do desejável (JAMES et al., 2013), o que inspirou o uso de várias árvores de decisão que agem em conjunto para gerar uma resposta. Essa técnica recebeu o nome de *Random Forest*.

Os conjuntos de árvores aleatórias são geradas aleatorizando tanto o subconjunto de dados que formará cada árvore com amostragem tipo *bootstrap* quanto um segundo processo de seleção aleatória para escolher apenas um subconjunto das covariáveis. Com todas as árvores formadas, a resposta é fornecida através de, em problemas de classificação, uma votação, em que cada árvore fornece uma resposta e a classe mais 'votada' se torna a resposta da floresta, ou de uma média, em problemas de regressão.

Por sua vez, as *Random Survival Forests* (ISHWARAN et al., 2008) são uma adaptação das *Random Forests* para acomodar dados com censura a direita. Os dados são separados objetivando folhas homogêneas, assim como nas árvores de decisão, para que um método não paramétrico retorne uma função de sobrevivência para cada nó folha. Então, uma média dos resultados é obtida como resposta final do modelo (SHAMSUTDINOVA et al., 2022).

A principal adaptação é realizada é no que diz respeito à divisão dos nós, já que, na RSF, os nós são divididos maximizando a diferença de sobrevivência entre os nós filhos ao invés de minimizar a variância como nas *random forests* normais. Isso é feito por meio da maximização do teste estatístico log-rank entre a função de risco cumulativo (NEWAZ; HAQ; AHMED, 2021). É o uso desse teste que permite o uso de dados censurados a direita no modelo, possibilitando sua aplicação na análise de sobrevivência (SNIDER; MCBEAN, 2021).

2.3.3 PL-kNN

Parameterless k*-Nearest Neighbors* (PL-kNN) (JODAS et al., 2022; JODAS et al., 2023) é uma variação do k*-Nearest Neighbors* tradicional feito com o objetivo de eliminar a otimização de hiperparâmetros. Isso ocorre porque, ao invés de ser informado um valor de k para determinar o número ótimo de vizinhos mais próximos, o modelo determina um valor ótimo de k com base no conjunto de dados. Isso é feito estabelecendo um conjunto de vizinhos relevantes por meio de uma região que estende da amostra de teste até o centróide mais próximo do seu cluster de classe em potencial, sendo que o número de vizinhos nessa região se tornam k e, portanto, as amostras dentro da região, os k vizinhos mais próximos. Algumas considerações são o uso da distância Manhattan para reduzir sensibilidade a escala, e a determinação de centróides usando a mediana ao invés da média, reduzindo o impacto de *outliers*. Outra medida contra *outliers* é a atribuição de pesos inversamente proporcionais à distância euclidiana dos seus respectivos centróides às amostras de treinamento.

Visto isso, o treinamento do algoritmo se resume à clusterização do conjunto de dados e atribuição de pesos às amostras de treinamento baseado nas suas distâncias aos centróides de suas classes. Durante a predição, é computada a distância da amostra de teste ao centróide mais próximo, a região semi-circular obtida, e os vizinhos dentro dela selecionados. Então, a classificação ocorre a partir de um voto ponderado, com amostras mais próximas recebendo peso maior na votação.

O processo de predição está definido a seguir (Eq. 20):

$$p_i(\boldsymbol{s}) = \frac{\sum\limits_{\forall \boldsymbol{z} \in \mathcal{T} \wedge \lambda(\boldsymbol{z}) = \omega_i} D_M(\boldsymbol{z}_j, \boldsymbol{s})^{-1} * W(\boldsymbol{z}_j, \boldsymbol{c}_i)}{\sum\limits_{k=1}^{n} p_k(\boldsymbol{s})}, \quad \forall i \in \mathcal{Y}, \tag{20}$$

onde $\mathcal{T} = \{\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_m\}$ é o conjunto de todas as amostras de treinamento dentro do semicírculo, $\lambda(\boldsymbol{z})$ representa o rótulo verdadeiro da amostra de treinamento $\boldsymbol{t}$, $D_M(\boldsymbol{z}_j, \boldsymbol{s})$ é a distância de Manhattan entre a amostra de teste $\boldsymbol{s}$ e seu j-ésimo vizinho mais próximo $\boldsymbol{z}_j$, $W(\boldsymbol{z}_j, c_i)$ é um peso representado pela distância inversa entre o j-ésimo vizinho mais próximo $\boldsymbol{z}_j$ e seu centróide c_i da classe i, $\mathcal{Y} = \{\omega_1, \omega_2, \omega_3, \ldots, \omega_n\}$ é o conjunto de classes do conjunto de dados, e $p_i(\boldsymbol{s})$ indica a probabilidade de a amostra $\boldsymbol{s}$ ser atribuída à classe ω_i.

Vale ressaltar que, nesse caso de uso, as ‘classes’ consideradas são a ocorrência do evento e censura, como foi comentado no início da seção 2.3 desse trabalho.

2.3.4 OPF-kNN

O *Optimum-Path Forest* (OPF) (PAPA; FALCAO; SUZUKI, 2009) é uma abordagem de classificação supervisionada baseada em grafos, projetada para superar limitações de classificadores tradicionais, como *Multilayer Perceptrons* (MLPs) e *Support Vector Machines* (SVMs). O modelo se destaca por não necessitar de ajuste fino de hiperparâmetros e por apresentar treinamento rápido mesmo em bases de dados extensas.

O OPF modela o conjunto de treinamento como um grafo, cujos nós representam amostras e cujas arestas representam relações de proximidade, ponderadas por uma métrica de distância no espaço de atributos. Na variante OPF-kNN, utilizada neste trabalho, o grafo é construído a partir da relação de k-vizinhos mais próximos, o que torna a estrutura mais robusta a variações no formato das classes.

Cada nó do grafo recebe um peso calculado por uma função de densidade de probabilidade (pdf) estimada a partir das distâncias para seus vizinhos, utilizando um kernel Gaussiano. Os máximos locais dessa função são escolhidos como *protótipos* ou *raízes* do modelo, servindo como pontos de origem para a expansão das árvores da *Optimum-Path Forest*. A pdf está definida na equação (21):

$$\rho(s) = \frac{1}{\sqrt{2\pi\sigma^2}\,k} \sum_{t \in A_k(s)} \exp\left(-\frac{d^2(s,t)}{2\sigma^2}\right), \qquad \sigma = \frac{d_f}{3}. \tag{21}$$

O algoritmo atribui a cada amostra um custo baseado no valor mínimo entre a densidade do nó e o valor de caminho ótimo do nó predecessor. O processo de crescimento das árvores é conduzido pelo *Image Foresting Transform* (IFT) (FALCAO et al., 1999), que propaga rótulos a partir dos protótipos para as demais amostras, criando regiões de influência que definem as partições do espaço de atributos.

Para garantir que os caminhos ótimos permaneçam dentro da classe correta, uma segunda função de custo é utilizada na etapa final do treinamento, atribuindo custo negativo ($-\infty$) para conexões entre amostras de classes diferentes. As funções de custo f_1 e f_2 estão definidas pelas equações (22) e (23):

$$f_1(\langle t \rangle) = \begin{cases} \rho(t), & t \in R, \\ \rho(t) - \delta, & \text{caso contrário}, \end{cases} \qquad f_1(\pi_s \cdot \langle s,t \rangle) = \min\{\, f_1(\pi_s),\, \rho(t) \,\}, \tag{22}$$

$$f_2(\langle t \rangle) = \begin{cases} \rho(t), & t \in R, \\ \rho(t) - \delta, & \text{caso contrário}, \end{cases} \qquad f_2(\pi_s \cdot \langle s,t \rangle) = \begin{cases} -\infty, & \lambda(t) \neq \lambda(s), \\ \min\{\, f_2(\pi_s),\, \rho(t) \,\}, & \text{caso contrário}. \end{cases} \tag{23}$$

O treinamento segue dois passos principais: primeiramente, se seleciona k^* testando valores de k no intervalo $[1, k_{max}]$ e escolhendo aquele que maximiza a acurácia no conjunto de treinamento. Então, é construída a floresta final, onde o IFT é executado novamente usando a função de custo restritiva, garantindo consistência de rótulos.

Para classificar uma nova amostra, calcula-se seu custo de conexão para todos os protótipos do conjunto de treinamento, atribuindo o rótulo do protótipo que oferece o caminho de menor custo. Esse processo preserva as propriedades estruturais do grafo e mantém a eficiência do método.

2.3.5 *Beran Estimator with Neural Kernels*

O método BENK (*Beran Estimator with Neural Kernels*) (KIRPICHENKO et al., 2024) é um modelo feito, originalmente, para estimar o *Conditional Average Treatment Effect* (CATE) em cenários com dados censurados de tempo até o evento, aplicando o estimador de Beran para obter funções de sobrevivência de grupos de controle e tratamento. Nesse trabalho, será aplicado como um Estimador de Beran comum, porém com uma importante diferença. A principal inovação é substituir os kernels convencionais do estimador de Beran por *neural kernels*, isto é, kernels implementados como redes neurais específicas e capazes de modelar de forma mais flexível estruturas complexas no espaço de atributos.

Como os kernels do BENK são aprendidos por redes neurais que compartilham parâmetros, é possível treinar o modelo mesmo em conjuntos de dados reduzidos. Além disso, o uso de *neural kernels* torna o método mais robusto a distribuições complexas dos dados, nas quais kernels fixos, como o Gaussiano, podem apresentar desempenho inferior.

A arquitetura do BENK é composta por múltiplos sub-redes idênticas que implementam os kernels neurais, normalização para cálculo dos pesos do estimador de Beran e o próprio cálculo da função de sobrevivência. Durante o treinamento, a perda utilizada é o *Median Squared Error* (MSE) entre o tempo de evento previsto e o tempo esperado, calculado a partir da função de sobrevivência estimada.

O método não requer grandes volumes de dados de treinamento devido ao seu esquema de amostragem eficiente, embora possua diversos hiperparâmetros que podem aumentar o tempo de treinamento. Limitações incluem a suposição de domínios semelhantes entre grupos e a necessidade de investigações futuras sobre versões robustas que lidem com outliers.

3 Metodologia

O presente trabalho propõe uma metodologia que cumpra três objetivos principais: o algoritmo desenvolvido deve minimizar a necessidade de otimização de hiperparâmetros, ao mesmo tempo que não dependa de um conhecimento prévio sobre o comportamento dos tempos do evento de interesse e acomode os efeitos da covariáveis na previsão.

3.1 Método proposto

O estimador de Beran pode ser representado através de uma função kernel $K(\mathbf{x}^*, \mathcal{V}, \mathcal{Y})$, onde $\mathbf{x}^*$ é o vetor de covariáveis alvo, $\mathcal{V}$ é o conjunto de amostras de treinamento, e $\mathcal{Y}$ representa os rótulos correspondentes indicando falha ou censura. Dentro desse framework, o estimador aprende a relação entre as covariáveis e as probabilidades de confiabilidade por meio de suavização baseada em kernel. No entanto, esse processo pode ser interpretado como um problema de classificação dentro do nosso paradigma. De forma geral, a função kernel $K(\mathbf{x}^*, \mathcal{V}, \mathcal{Y})$ computa um peso para cada observação de treinamento em relação com a covariável alvo $\mathbf{x}^*$. Esse peso é influenciado por fatores diferentes para cada modelo, com a principal semelhança sendo a influência dos rótulos $\mathcal{Y}$, que fornecem informações cruciais sobre o status de sobrevivência ou censura da observação.

O Algoritmo 1 implementa de forma geral o estimador de Beran proposto com as novas funções kernel. O algoritmo estima a função de confiabilidade utilizando o kernel proposto baseado no PL-kNN e no OPF-kNN. O procedimento opera em dois principais loops: (i) o primeiro calcula os pesos do kernel para cada ponto no tempo, e (ii) o segundo estima a probabilidade de confiabilidade em cada passo do tempo. O algoritmo começa inicializando as variáveis necessárias nas Linhas 1-3. O total de observações é armazenado em n (Linha 1). Um array inicializado com zero, S, armazena as probabilidades de confiabilidade (Linha 2). Um array vazio, M, também é preparado para acumular os pesos de kernel calculados para uso posterior (Linha 3).

No primeiro loop (Linha 4), para cada ponto de tempo $T[i]$, o algoritmo identifica o conjunto de observações ainda em risco, definindo-as como aquelas cujos eventos satisfazem $T[j] \geq T[i]$ (Linha 5). As covariáveis correspondentes são extraídas de $\mathcal{V}$ e os indicadores de censura dos indivíduos em risco são extraídos de $\mathcal{Y}$ nas Linhas 6 e 7, respectivamente. Um kernel K é aplicado às covariáveis alvo $\mathbf{x}^*$ e ao conjunto de treinamento ainda em risco, $\mathcal{V}$, para computar um peso representando a influência local de cada observação. O par $(\mathcal{V}, \mathcal{Y})$ é utilizado para treinar o os modelos-base a fim

de computar os pesos para o novo ponto de vista $\mathbf{x}^*$.

Algorithm 1 Estimativa da Função de Confiabilidade Baseada em Kernel.

Require: Lista ordenada de tempos T, conjunto de treinamento $\mathcal{X}$, conjunto de indicações de falha ou censura Δ, uma amostra de teste $\mathbf{x}^*$, e uma função kernel K

Require: variáveis auxiliares: lista de pesos M, lista de probabilidades de confiabilidade S

1: $n \leftarrow |T|$
2: $S \leftarrow \text{zeros}(n)$
3: $M \leftarrow [\,]$
4: **for** $i = 0$ até $n - 1$ **do**
5: $\quad \mathcal{R}_i \leftarrow \{j \mid T[j] \geq T[i]\}$
6: $\quad \mathcal{V} \leftarrow \{\mathcal{X}[j] \mid j \in \mathcal{R}_i\}$
7: $\quad \mathcal{Y} \leftarrow \{\Delta[j] \mid j \in \mathcal{R}_i\}$
8: $\quad w_i \leftarrow K(\mathbf{x}^*, \mathcal{V}, \mathcal{Y})$
9: $\quad M \leftarrow M \,\|\, w_i$
10: **end for**
11: **for** $i = 0$ até $n - 1$ **do**
12: $\quad \mathcal{R}_i \leftarrow \{j \mid T[j] \geq T[i]\}$
13: $\quad P \leftarrow \{M[j] \mid j \in \mathcal{R}_i\}$
14: $\quad \psi_i \leftarrow \dfrac{\sum_{k \in \mathcal{R}_i} P[k] \cdot \Delta[k]}{\sum_{k \in \mathcal{R}_i} P[k] + \epsilon}$
15: $\quad$ **if** $i = 0$ **then**
16: $\quad\quad S_i \leftarrow 1 - \psi_i$
17: $\quad$ **else**
18: $\quad\quad S_i \leftarrow S_{i-1} \cdot (1 - \psi_i)$
19: $\quad$ **end if**
20: **end for**
21: **return** S

O segundo loop itera através de cada instante de tempo $T[i]$ mais uma vez para estimar a função de confiabilidade (Linha 11). O conjunto de observações em risco é identificado (Linha 12) juntamente com os pesos correspondentes (Linha 13). Uma soma ponderada das observações em risco é computada utilizando os pesos do kernel (Linha 14). A função de risco para cada ponto de tempo i é então calculada da seguinte forma (Eq. 24):

$$\psi_i = \frac{\mathcal{P}[i] \cdot \Delta_i}{\sum_{k=0}^{n-1} \mathcal{P}[k] + \epsilon}, \tag{24}$$

assegurando estabilidade numérica com uma constante pequena $\epsilon = 10^{-10}$ no denominador. A probabilidade de confiabilidade no primeiro passo temporal é definida como $S_0 = 1 - \psi_0$. Para os passos subsequentes, a probabilidade de sobrevivência é computada recursivamente como (Eq. 25):

$$S_i = S_{i-1}(1 - \psi_i). \tag{25}$$

Essa abordagem recursiva segue o mesmo princípio da estimação de Kaplan-Meier, mas incorpora ponderação baseada em kernel para adaptar-se às informações das covariáveis. Finalmente, os valores computados da função de confiabilidade, representados por S, são retornados como saída do algoritmo. Essa função estima a probabilidade de sobrevivência ao longo do tempo, levando em consideração os efeitos das covariáveis através da suavização por kernel.

Assim, propõe-se três funções kernel para serem usada nessa abordagem, uma dada pela média dos pesos dos vizinhos mais próximos calculados pelo PL-kNN, outra baseada na média dos custos dos nós mais próximos calculados pelo OPF-kNN e uma última correspondente à média dos pesos das arestas calculados por esse mesmo modelo.

Os algoritmos 2, 3 e 4 calculam os valores dessas funções. Todos operam extraíndo informações calculadas pelos seus modelos base, assim necessitando que o modelo seja treinado com o conjunto de treinamento $\mathcal{X}$ e que uma previsão seja feita com a covariável alvo atual $\mathbf{x}^*$. No algoritmo 2, após esse passo, basta juntar os pesos individuais, pertencentes ao conjunto de pesos calculados pelo modelo, e calcular a média desses.

Algorithm 2 Função de kernel PL-kNN

Require: Uma amostra de teste $\mathbf{x}^*$, o conjunto de treinamento $\mathcal{X}$ e o conjunto de rótulos $\mathcal{Y}$
Require: variáveis auxiliares: lista de pesos p
1: **function** $AdquirePesosPL - kNN(\mathbf{x}^*, \mathcal{X}, \mathcal{Y})$
2: $\quad Treino(PL - kNN, \mathcal{X}, \mathcal{Y})$
3: $\quad Previsão(PL - kNN, \mathbf{x}^*)$
4: $\quad p \leftarrow \{w | w \in W(z_j, c_i)\}$
5: $\quad$ **return** p
6: **end function**
7: $W \leftarrow AdquirePesosPL - kNN(\mathbf{x}^*, \mathcal{X}, \mathcal{Y})$
8: $K \leftarrow \frac{\sum_{k=0}^{|W|-1} W_k}{|W| + \epsilon}$
9: **return** $\frac{K}{n+\epsilon}$

Os pesos são extraídos dentro da função definida nas Linhas 1-6. Finalizada a previsão (Linha 3), os pesos individuais dos vizinhos mais próximos, dados por w e pertencentes ao conjunto de pesos $W(z_j, c_i)$, são inseridos na lista p (Linha 4) e, então, essa lista se torna o valor de retorno da função. Tendo adquirido esse valor, a média é calculada (Linha 8). Essa média é dada por (Eq. 26):

$$K \leftarrow \frac{\sum_{k=0}^{|W|-1} W_k}{|W| + \epsilon} \tag{26}$$

assegurada a estabilidade numérica pela pequena constante $\epsilon = 10^{-6}$. Tem-se, por fim, o valor final da função de kernel.

Algorithm 3 Função de kernel OPF-kNN baseado nos custos dos nós vizinhos

Require: Uma amostra de teste $\mathbf{x}^*$, o conjunto de treinamento $\mathcal{X}$ e o conjunto de rótulos $\mathcal{Y}$
Require: variáveis auxiliares: lista de custos C, tamanho de teste de treino α e semente de aleatorização $seed$
1: **function** $AdquireCustosOPF - kNN(\mathbf{x}^*, \mathcal{X}, \mathcal{Y})$
2: $C \leftarrow [\,]$
3: **if** $|\mathcal{X}| > 2$ **then**
4: $\alpha \leftarrow 0.25$
5: $seed \leftarrow 1$
6: $\mathcal{X}^*, \mathcal{X}', \mathcal{Y}^*, \mathcal{Y}' \leftarrow TrainTestSplit(\mathcal{X}, \mathcal{Y}, \alpha, seed)$
7: $Treino(OPF - kNN, \mathcal{X}^*, \mathcal{X}', \mathcal{Y}^*, \mathcal{Y}')$
8: $Previsão(OPF - kNN, \mathbf{x}^*)$
9: **for** $k = 0$ to $BestK$ **do**
10: $C \parallel \rho(k)$
11: **end for**
12: **return** C
13: **end if**
14: **end function**
15: $C \leftarrow AdquireCustosOPF - kNN(\mathbf{x}^*, \mathcal{X}, \mathcal{Y})$
16: $K \leftarrow \frac{\sum_{k=0}^{|C|-1} C_k}{|C| + \epsilon}$
17: **return** K

No kernel baseado no OPF-kNN, que opera conforme o algoritmo 3, a implementação do OPF utilizada nesse trabalho, OPFython, espera uma divisão do conjunto de dados entre conjunto de teste e conjunto validação. Para tal fim, empregou-se a função '*TrainTestSplit*' da biblioteca *scikit-learn*, que, por sua vez, espera dois parâmetros de entrada além do conjunto de covariáveis e de variáveis alvo, sendo essas o tamanho do conjunto de teste (ou validação), representado por α definido na Linha 4, e uma semente de aleatorização usada no embaralhamento das amostras do conjunto, representada por $seed$ e definida na linha 5. Com isso, obtemos $\mathcal{X}^*$, as covariáveis do conjunto de treino, $\mathcal{Y}^*$, os rótulos (censura e falha) do conjunto de treino, $\mathcal{X}'$, as covariáveis do conjunto validação, e $\mathcal{Y}'$, os rótulos do conjunto validação.

Além disso, foi determinado durante os experimentos que a implementação falhava quando o conjunto $\mathcal{X}$ tinha apenas uma amostra, algo raro em casos de uso comuns, mas garantido quando agindo como um kernel no Estimador de Beran, sendo

que o estimador considera uma amostra a menos a cada iteração. Tendo isso em vista, é necessário desconsiderar a última iteração do algoritmo de Beran para garantir a execução do modelo proposto.

Aqui, os custos que determinam o peso do kernel são as densidades atribuídas a cada amostra pelo algoritmo do OPF-kNN dadas pela função de densidade probabilística e representada por $\rho(k)$. A média dos custos dos k vizinhos mais próximos é determinada e retornada como peso do kernel, sendo que o valor ótimo de k, dado por $BestK$, é determinado ao decorrer do treino do OPF, na linha 7. Esse valor deve aderir a um intervalo de 1 até um máximo determinado na construção da classe, conforme a implementação na biblioteca usada.

Algorithm 4 Função de kernel OPF-kNN baseado nos pesos das arestas dos nós vizinhos

Require: Uma amostra de teste $\mathbf{x}^*$, o conjunto de treinamento $\mathcal{X}$ e o conjunto de rótulos $\mathcal{Y}$

Require: variáveis auxiliares: lista de pesos das arestas D, tamanho de teste de treino α e semente de aleatorização $seed$

1: **function** $AdquirePesosArestasOPF - kNN(\mathbf{x}^*, \mathcal{X}, \mathcal{Y})$
2: $C \leftarrow [\,]$
3: **if** $|\mathcal{X}| > 2$ **then**
4: $\alpha \leftarrow 0.25$
5: $seed \leftarrow 1$
6: $\mathcal{X}^*, \mathcal{X}', \mathcal{Y}^*, \mathcal{Y}' \leftarrow TrainTestSplit(\mathcal{X}, \mathcal{Y}, \alpha, seed)$
7: $Treino(OPF - kNN, \mathcal{X}^*, \mathcal{X}', \mathcal{Y}^*, \mathcal{Y}')$
8: $Previsão(OPF - kNN, \mathbf{x}^*)$
9: **for** $k = 0$ to $BestK$ **do**
10: $D \parallel d(\mathbf{x}^*, k)$
11: **end for**
12: **return** D
13: **end if**
14: **end function**
15: $D \leftarrow AdquirePesosArestasOPF - kNN(\mathbf{x}^*, \mathcal{X}, \mathcal{Y})$
16: $K \leftarrow \dfrac{\sum_{k=0}^{|C|-1} C_k}{|C| + \epsilon}$
17: **return** K

No algoritmo 4, o algoritmo se baseia nas arestas da floresta ao invés de seus nós, com o valor extraído do OPF se tratando das distâncias entre os nós, dada pela função de valor de caminho (*path-value function*) representada por $d(\mathbf{x}^*, k)$.

3.2 Conjuntos de dados

Para o treinamento e validação dos modelos propostos foram empregados três conjuntos de dados referentes tanto à aplicações médicas quanto da engenharia. São

esses:

1. Válvulas de segurança de subsuperfície (DHSV) (COLOMBO, 2023): um conjunto de dados relacionado à avaliação de válvulas DHSV, incluindo 576 amostras coletadas de poços de petróleo.

2. *Survival analysis of heart failure patients: A case study*[1] (AHMAD et al., 2017): um conjunto de dados de análise de sobrevivência em insuficiência cardíaca, incluindo 299 amostras de pacientes acompanhados com variáveis clínicas e de laboratório.

3. *Survival analysis data for metastatic Prostate cancer patients at UCI*[2] (KATONGOLE; SANDE; NIYONZIMA, 2021): Um conjunto de dados de análise de sobrevivência para pacientes com câncer de próstata metastático atendidos no UCI, incluindo 160 amostras com desfecho e tempo até o evento em meses.

Cada uma das amostras é definida por um conjunto de covariáveis, definidas para cada conjunto nas tabelas 3, 4 e 5:

Tabela 3 – Descrição das covariáveis que representam as amostras de dados do DHSV.

Covariável	**Descrição**
Water Column (WC)	Profundidade da coluna d'água em que o DHSV está instalado no poço.
Manufacturer	Empresa responsável pela fabricação do DHSV.
Censoring indication	Variável binária indicando se a amostra é censurada (evento = 0) ou falha (evento = 1).
Time to event	Tempo registrado até ocorrer uma falha ou até a observação ser censurada.

Fonte: Elaborada pelo autor.

[1] Disponível para *download* em: <https://plos.figshare.com/articles/dataset/Survival_analysis_of_heart_failure_patients_A_case_study/5227684/1?file=8937223>. Acesso em: 17/11/2025

[2] Disponível para *download* em: <https://figshare.com/articles/dataset/Survival_analysis_data_for_metastatic_Prostate_cancer_patients_at_UCI/13625627?file=26150744>. Acesso em: 17/11/2025

Tabela 4 – Descrição das covariáveis que representam as amostras do conjunto *Survival analysis of heart failure patients: A case study*.

Covariável	**Descrição**
Age	Idade do paciente (anos).
Gender	Sexo do paciente (0 = feminino, 1 = masculino).
Anaemia	Indicador de anemia (0 = não, 1 = sim).
Diabetes	Indicador de diabetes (0 = não, 1 = sim).
BP	Hipertensão arterial (*high blood pressure*; 0 = não, 1 = sim).
Smoking	Hábito de fumar (0 = não, 1 = sim).
Ejection fraction	Fração de ejeção do ventrículo esquerdo (%).
Sodium	Sódio sérico (mEq/L).
Creatinine	Creatinina sérica (mg/dL).
CPK	Creatina fosfoquinase (U/L).
Pletelets	Plaquetas (kiloplaquetas/mL).
TIME	Tempo de acompanhamento até o desfecho ou censura (dias).

Fonte: Elaborada pelo autor.

Tabela 5 – Descrição das covariáveis que representam as amostras do conjunto *Survival analysis data for metastatic Prostate cancer patients at UCI*.

Covariável	**Descrição**
Age	Idade do paciente (anos) no momento da apresentação inicial.
Presentation date	Data de apresentação/primeira consulta do caso.
Baseline PSA	PSA basal na apresentação (ng/mL).
Gleason Score	Escore de Gleason do tumor prostático na biópsia.
CNS	Metástase em sistema nervoso central (0 = ausente, 1 = presente).
LIVER	Metástase hepática (0 = ausente, 1 = presente).
LUNG	Metástase pulmonar (0 = ausente, 1 = presente).
BONES	Metástase óssea (0 = ausente, 1 = presente).
SPLEEN	Metástase esplênica (0 = ausente, 1 = presente).
KIDNEY	Metástase renal (0 = ausente, 1 = presente).
BONE MARROW	Metástase em medula óssea (0 = ausente, 1 = presente).
STATUS	Situação no último acompanhamento (*Alive* / *Death*).
Patient time of update	Data/tempo da última atualização cadastral do paciente.[3]
Time to event -Months	Tempo até o evento (óbito) ou censura, em meses.

Fonte: Elaborada pelo autor.

3.3 Configuração dos experimentos

Para cada conjunto de dados, as informações exceto o indicador de ocorrência de evento e o tempo até o evento foram usadas como covariáveis nos modelos preditivos. Visto que os modelos propostos são baseados em algoritmos baseados em distâncias, como o *k-NN*, usou-se da normalização dos dados por meio do algoritmo *z-score*, o qual transforma uma variável numérica para ter média 0 e desvio padrão 1, colocando-as na mesma escala. Por fim, variáveis categóricas, isto é, covariáveis não-numéricas como, por exemplo, o fabricante do equipamento DHSV, foram codificadas utilizando *One Hot Encoder*, que converte uma categoria nominal em várias colunas binárias (0/1), uma por categoria.

Os novos modelos foram avaliado o comparando com quatro modelos *baseline*: o estimador de Kaplan-Meier, modelo proporcional de Cox (Cox PHM), *Random Survival Forests* (RSF), e Estimador de Beran com Kernels Neurais (BENK).

Embora os dois primeiros modelos sirvam como *benchmarks* amplamente utilizados em análise de sobrevivência, o foco principal de nossa comparação é o estimador BENK, devido à sua configuração abrangente de hiperparâmetros para modelar padrões de sobrevivência complexos. O modelo BENK é composto, principalmente, por um conjunto de redes neurais que recebem como entrada um par de covariáveis de amostras.

O BENK admite dois hiperparâmetros para a seleção de subconjuntos de amostras: (i) uma fração do conjunto de dados completo (n) que compõe a entrada das redes neurais; e (ii) o número de repetições de seleção aleatória desse subconjunto (N).

Dada a ampla calibração de hiperparâmetros exigida pelo BENK, foram testados 10 valores escolhidos aleatoriamente para N e n, avaliando exaustivamente seu comportamento na geração da função de sobrevivência. Adaptou-se o código original do BENK disponível no GitHub[4].

Duas particularidades da configuração vigente estão na seleção de hiperparâmetros do *Random Survival Forests* e no tratamento do conjunto de dados referente a manutenção preditiva. No primeiro, foi empregado o algoritmo de otimização de hiperparâmetros *Randomized Search*, que faz amostragens aleatórias de combinações de hiperparâmetros a partir de determinadas distribuições (ou listas), avaliando cada combinação via *cross-validation* e escolhendo a melhor. A escolha do algoritmo foi influenciada pelas restrições de tempo para o desenvolvimento do trabalho, visto que a realização dos experimentos já consome uma grande quantidade de tempo pela todas as combinações de hiperparêmtros em conjuntos amplos de dados, um aspecto inerente do algoritmo *Grid Search*, tornando de bom tom economizar tempo na otimização

[4] <https://github.com/Stasychbr/BENK>. Acesso em: 17/11/2025

de hiperparâmetros dos modelos de comparação.

3.4 Métricas de Avaliação

Para avaliar o desempenho dos modelos, três métricas foram empregadas: o *Integrated Brier Score* (IBS) (GRAF et al., 1999), voltado a capturar a capacidade discriminante do modelo em termos de acurácia; o *Concordance Index* (C-Index) (JR; LEE; MARK, 1996); e sua variante com ponderação pela Probabilidade Inversa de Censura (*C-Index IPCW*) (UNO et al., 2011), ambas utilizadas para mensurar a habilidade de ranqueamento do modelo. Essas métricas são bem conhecidas e amplamente utilizadas na literatura para a avaliação de modelos baseados em análise de confiabilidade.

3.4.1 Integrated Brier Score

O Integrated Brier Score (IBS) é uma métrica de avaliação da capacidade de predição em modelos de análise de sobrevivência. Esta métrica é similar ao erro quadrático médio entre as probabilidades de sobrevivência previstas pelo modelo e a indicação do evento observado ao longo de um intervalo de tempo de interesse. Formalmente, o Brier Score no tempo t é definido pela equação (27) como (PRINCE et al., 2025):

$$\text{BS}(t) = \frac{1}{n}\sum_{i=1}^{n}\left[\hat{S}(t|X_i) - I(T_i > t)\right]^2, \tag{27}$$

onde n é o número de observações, $\hat{S}(t|X_i)$ representa a probabilidade de sobrevivência estimada para o indivíduo i condicionada às suas covariáveis $\boldsymbol{X}_i$, T_i denota o tempo observado, e $I(T_i > t)$ assume o valor 1 se o indivíduo sobreviveu além do tempo t e 0 caso contrário. O IBS é calculado como (Eq 28):

$$\text{IBS} = \frac{1}{t_{max}}\int_0^{t_{max}} \text{BS}(t)\, dt, \tag{28}$$

cujo valor representa a média ponderada dos valores de Brier Scores ao longo do intervalo $[0, t_{max}]$.

3.4.2 Concordance Index

O Concordance Index (C-Index), estima a capacidade de um modelo de sobrevivência através da proporção de pares concordantes entre todos os pares comparáveis

de observações. Formalmente, o C-Index de Harrell é representado de acordo com a equação (29) (UNO et al., 2011):

$$\text{C-Index} = \frac{\sum_{i \neq j} 1_{T_i < T_j} \cdot 1_{\eta_i > \eta_j} \cdot \delta_i}{\sum_{i \neq j} 1_{T_i < T_j} \cdot \delta_i}, \tag{29}$$

onde T_i e T_j representam os tempos observados para os indivíduos i e j, η_i e η_j são as probabilidades de risco estimadas pelo modelo, δ_i é o indicador de evento ($\delta_i = 1$ se o evento foi observado, $\delta_i = 0$ se censurado), e $1(\cdot)$ é a função indicadora que retorna 1 se a condição for satisfeita, ou zero caso contrário. Um par (i, j) é considerado comparável quando $T_i < T_j$ e o evento foi observado para o indivíduo i ($\delta_i = 1$), e concordante quando o modelo atribui maior risco ao indivíduo que experimentou o evento primeiro ($\eta_i > \eta_j$). O índice varia entre 0 e 1, onde 0.5 representa discriminação equivalente ao acaso, valores superiores a 0.7 são geralmente considerados indicativos de boa capacidade discriminatória, e 1.0 representa concordância perfeita.

3.4.3 C-Index com Ponderação pela Probabilidade Inversa de Censura

O C-Index com ponderação pela probabilidade inversa de censura (C-Index IPCW) representa uma melhoria do C-Index tradicional para reduzir o viés introduzido pela censura através da incorporação de pesos baseados na distribuição temporal da censura. A formulação do C-Index IPCW é dada pela equação (30) (PRINCE et al., 2025):

$$\omega_i(t) = \frac{I[T_i^* \leq t, \delta_i = 1]}{\hat{G}_i(T_i^-)} + \frac{I[T_i^* > t]}{\hat{G}_i(t)}, \tag{30}$$

onde $\hat{G}_i(t)$ é a probabilidade de não ser censurado até o tempo t, $\hat{G}_i(T_i^-)$ é a probabilidade de não ser censurado imediatamente antes do tempo observado T_i, e $\frac{I[T_i^* > t]}{\hat{G}_i(t)}$ representa o peso para indivíduos que ainda estão em risco no tempo t.

4 Experimentos e discussão

4.1 Análise quantitativa

As tabelas 6, 7 e 8 apresentam os resultados de cada modelo de referência usados para comparação e as mesmas métricas de avaliação usadas na metodologia proposta aplicados aos conjuntos de dados de válvulas DHSV, pacientes de insuficiência cardíaca e pacientes de câncer de próstata metastático, respectivamente. Os melhores resultados para cada métrica estão destacados em negrito.

Tabela 6 – Métricas de avaliação obtidas das amostras de teste do conjunto de válvulas DHSV.

Estimador	**IBS**	**C-index**	**C-index IPCW**
Cox PHM	0.345	0.650	0.406
RSF	0.500	0.582	0.477
Beran PL-kNN	0.323	0.566	0.525
Beran OPF-kNN	**0.256**	**0.773**	0.586
Beran OPF-kNN (Arc weights)	0.308	0.561	0.567
BENK N=1/n=28	0.306	0.525	0.493
BENK N=9/n=69	0.390	0.565	0.556
BENK N=10/n=49	0.398	0.499	0.557
BENK N=13/n=55	0.405	0.527	0.585
BENK N=14/n=85	0.436	0.528	**0.591**
BENK N=20/n=9	0.352	0.711	0.590
BENK N=25/n=65	0.528	0.470	0.418
BENK N=26/n=30	0.475	0.452	0.434
BENK N=35/n=54	0.502	0.417	0.368
BENK N=39/n=74	0.505	0.458	0.332

Fonte: Elaborado pelo autor

Entre os métodos propostos, se destaca o ótimo desempenho do Beran OPF-kNN, o qual obteve o melhor resultado em termos tanto do IBS (0.256) quanto do C-index (0.773), ficando abaixo apenas da configuração BENK com hiperparâmetros $N = 14$ e $n = 85$, em termos do C-index IPCW (0.591 para esta configuração do BENK em comparação com o valor 0.586 obtido pelo Beran OPF-kNN). Entretanto, a diferença é ínfima para este caso. A boa performance nessa terceira métrica, em particular, o mostra pouco suscetível à alta taxa de censura do conjunto de dados vigente, característica compartilhada entre os três modelos propostos, apesar do desempenho inferior dos outros modelos propostos: Beran PL-kNN e Beran OPF-kNN (*Arc weights*). Isso se torna especialmente valioso ao observar o efeito negativo que a

alta taxa de censura teve, por exemplo, no RSF.

O Beran PL-kNN e o Beran OPF-kNN que utiliza os pesos das arestas do subgrafo ("Beran OPF-kNN (*Arc weights*)"), apesar de obterem resultados inferiores em comparação com o Beran OPF-kNN regular, ainda demonstraram performance superior em comparação com os modelos de referência, superando o Cox em termos do IBS e do C-index IPCW e tendo desempenho comparável ou superior à diversas configurações do BENK.

Em termos do Beran PL-kNN, apenas o BENK com $N = 1$ e $n = 28$ obteve IBS menor (0.306) e apenas o BENK com $N = 20$ e $n = 9$ obteve C-index maior (0.771), porém, essa configuração foi superior nessa métrica por uma margem mais significativa. O C-index IPCW (0.525) do novo modelo foi o menor entre os métodos propostos, mas melhor que metade das configurações do BENK apresentadas.

O Beran OPF-kNN (*Arc weights*), por sua vez, praticamente iguala, no IBS (0.308), o BENK com configuração $N = 1$ e $n = 28$, que supera o Beran PL-kNN nessa métrica, mas é superado por duas configurações no seu C-index (0.561) ao invés de apenas uma. Seu C-index IPCW (0.567), apesar de marginalmente maior, fica atrás de apenas duas configurações do Beran com *kernels* neurais.

A tabela 7 apresenta os resultados para o conjunto de dados de insuficiência cardíaca.

Tabela 7 – Métricas de avaliação obtidas das amostras de teste do conjunto de insuficiência cardíaca.

Estimador	**IBS**	**C-index**	**C-index IPCW**
Cox PHM	**0.178**	0.632	0.662
RSF	0.184	**0.653**	**0.709**
Beran PL-kNN	0.245	0.417	0.374
Beran OPF-kNN	0.251	0.412	0.529
Beran OPF-kNN (Arc weights)	0.247	0.522	0.519
BENK N=1/n=28	0.219	0.519	0.606
BENK N=9/n=69	0.253	0.418	0.537
BENK N=10/n=49	0.262	0.483	0.583
BENK N=13/n=55	0.270	0.462	0.537
BENK N=14/n=85	0.271	0.522	0.591
BENK N=20/n=9	0.256	0.368	0.411
BENK N=25/n=65	0.247	0.524	0.501
BENK N=26/n=30	0.252	0.488	0.457
BENK N=35/n=54	0.292	0.453	0.424
BENK N=39/n=74	0.274	0.499	0.453

Fonte: Elaborado pelo autor

Analisando os resultados da tabela 7, observa-se a forma como a taxa reduzida

de censura desse conjunto de dados, que tem a menor taxa entre os três utilizados, facilita o funcionamento do RSF, que obteve os melhores valores em ambas as métricas C-index (0.653 para o C-index e 0.709 para o C-index IPCW) e praticamente iguala o melhor IBS obtido pelo modelo de Cox, com uma diferença ínfima de 0.006 entre ambos. Por outro lado, a alta dimensionalidade do conjunto, com amostras de onze covariáveis cada, perceptivelmente afeta os novos modelos propostos, que são todos baseados em distâncias.

Sendo assim, os três métodos propostos mostraram resultados extremamente similares e medianos no IBS, superando a maioria das configurações do BENK, porém, marginalmente. O Beran OPF-kNN baseado nas arestas foi melhor em termos do C-index (0.522), medida na qual foi comparável ou melhor que todas as configurações do BENK, mas, quando levando em conta a censura, igualou-se à sua contrapartida baseada nos pesos dos nós. Nessa terceira medida, o Beran PL-kNN obteve resultados inferiores comparado com os modelos testados, enquanto os demais modelos propostos demonstraram desempenho mediano.

Apesar da proximidade em performance do BENK, os novos modelos ainda produziram resultados inferiores para algumas métricas, tanto por um modelo de aprendizado de máquina na forma do RSF, quanto às obtidas pelo modelo tradicional estatístico: de Cox. Em relação ao primeiro, esses modelos ainda possuem a vantagem de tempo de execução ao minimizar ou descartar a otimização de hiperparâmetros. Entretanto, o modelo de Cox obteve desempenho superior nesse cenário, mesmo não necessitando de otimização de hiperparâmetros.

Um terceiro panorama pode ser observado nos resultados advindos dos dados de câncer de próstata (tabela 8). Esse conjunto de dados, assim como o primeiro analisado, apresenta alta taxa de censura, levando a um desempenho significativamente inferior do que o anterior para o RSF. Nesse teste, destacou-se o BENK com N = 35 e n = 54, obtendo o melhor IBS (0.187) e, por uma margem considerável, o melhor C-index (0.603), mostrando-se o modelo mais robusto para essas amostras.

Tabela 8 – Métricas de avaliação obtidas das amostras de teste do conjunto de cancer de próstata.

Estimador	**IBS**	**C-index**	**C-index IPCW**
Cox PHM	0.211	0.241	**0.742**
RSF	0.199	0.414	0.323
Beran PL-kNN	0.190	0.500	0.290
Beran OPF-kNN	0.207	0.345	0.613
Beran OPF-kNN (Arc weights)	0.212	0.517	0.129
BENK N=1/n=28	0.220	0.345	0.645
BENK N=9/n=69	0.331	0.241	0.097
BENK N=10/n=49	0.350	0.190	0.032
BENK N=13/n=55	0.321	0.241	0.032
BENK N=14/n=85	0.308	0.207	0.129
BENK N=20/n=9	0.281	0.172	0.065
BENK N=25/n=65	0.347	0.310	0.323
BENK N=26/n=30	0.242	0.328	0.452
BENK N=35/n=54	**0.187**	**0.603**	0.452
BENK N=39/n=74	0.219	0.448	0.677

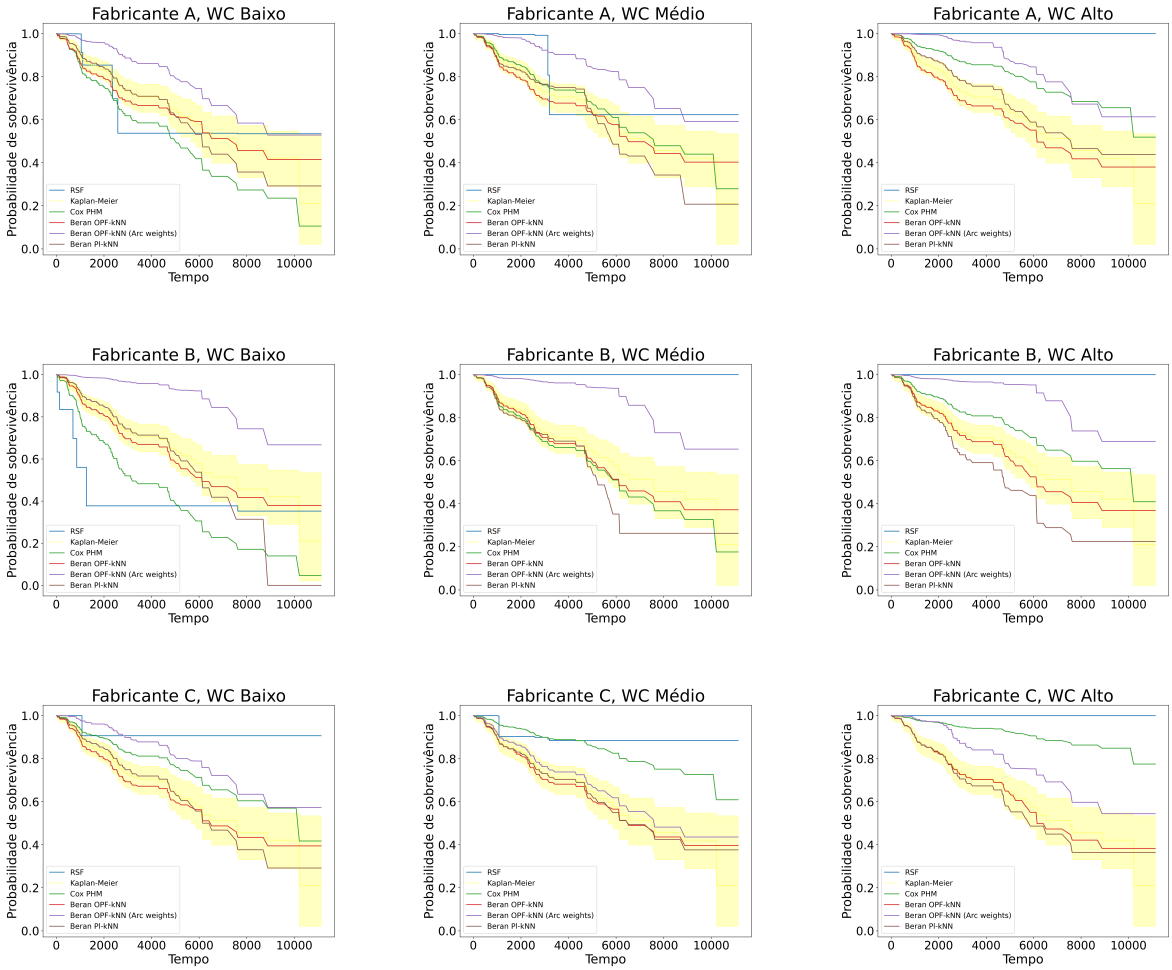
Fonte: Elaborado pelo autor

Os resultados dos métodos propostos indicam que esse conjunto de dados possui censura informativa, isto é, observações que não foram aleatoriamente censuradas no conjunto. Isso pode ocorrer quando a censura está relacionada ao risco do paciente - por exemplo, pacientes de alto risco podem abandonar o estudo devido a efeitos colaterais do tratamento, mudar para terapias alternativas, ou serem censurados por outras razões relacionadas à sua condição. Neste caso, os resultados indicam que o arc weights tem bom desempenho aparente (alto C-Index), mas esse desempenho não considera padrões de censura com padrões de sobrevivência verdadeiros. Quando o C-Index IPCW corrige para a censura informativa, o desempenho real se revela baixo, apresentando um valor de 0.129 para o C-index IPCW.

Por outro lado, o OPF-kNN mantém bom desempenho mesmo após a correção C-index IPCW, indicando que ele consegue distinguir verdadeiramente entre pacientes de alto e baixo risco, independentemente dos padrões de censura, se saindo melhor nesse quesito quando comparado com a melhor configuração do BENK. Neste caso, o OPF-kNN produziu um C-index IPCW de 0.613, enquanto a melhor configuração do BENK produziu um valor de 0.452 para a mesma métrica.

Outros fatores que podem ter prejudicado o desempenho do Beran PL-kNN e do Beran OPF-kNN Arc são a presença de dados de alta dimensionalidade e baixo número de amostras, hipótese essa que seria apoiada pelos resultados do último conjunto analisado, de dados de insuficiência cardíaca, que compartilhava essas características e também produziu métricas abaixo do esperado.

Figura 2 – Funções de sobrevivência para cada combinação de fabricante e profundidade operacional do conjunto de dados DHSV.



Fonte: Imagens elaboradas pelo autor

4.2 Análise qualitativa

O desempenho dos modelos propostos foi analisado qualitativamente ao comparar suas curvas de sobrevivência com as produzidas pelos modelos de referência e pelo estimador de Kaplan-Meier. Para os dados de válvula de segurança de subsuperfície, o efeito da covariável Coluna de Água (WC) foi observado estratificando-a em grupos baseados em valores baixos, médios e altos, seguindo orientações de especialistas na área de confiabilidade de poços acerca do histórico de profundidade de operação das válvulas. Para os demais conjuntos de dados, devido ao alto número de covariáveis e observações, foram selecionadas aleatoriamente nove amostras e estimadas as suas funções de sobrevivência com cada um dos modelos propostos, cada um dos modelos de referência e o estimador de Kaplan-Meier. As figuras 2, 3 e 4 mostram as curvas de sobrevivência obtidas para os conjuntos de válvulas DHSV, insuficiência cardíaca e câncer de próstata, respectivamente.

As curvas do estimador de Kaplan-Meier podem exibir planícies de probabilidades constantes devido a dados censurados. Na figura 2, o efeito da censura no RSF pode novamente ser constatado observando essas mesmas planícies em suas predições, levando a quedas bruscas após longos períodos de constância e, em combinações especialmente raras entre as amostras, até funções com probabilidade constante de 1 ao longo de todo o tempo de observação.

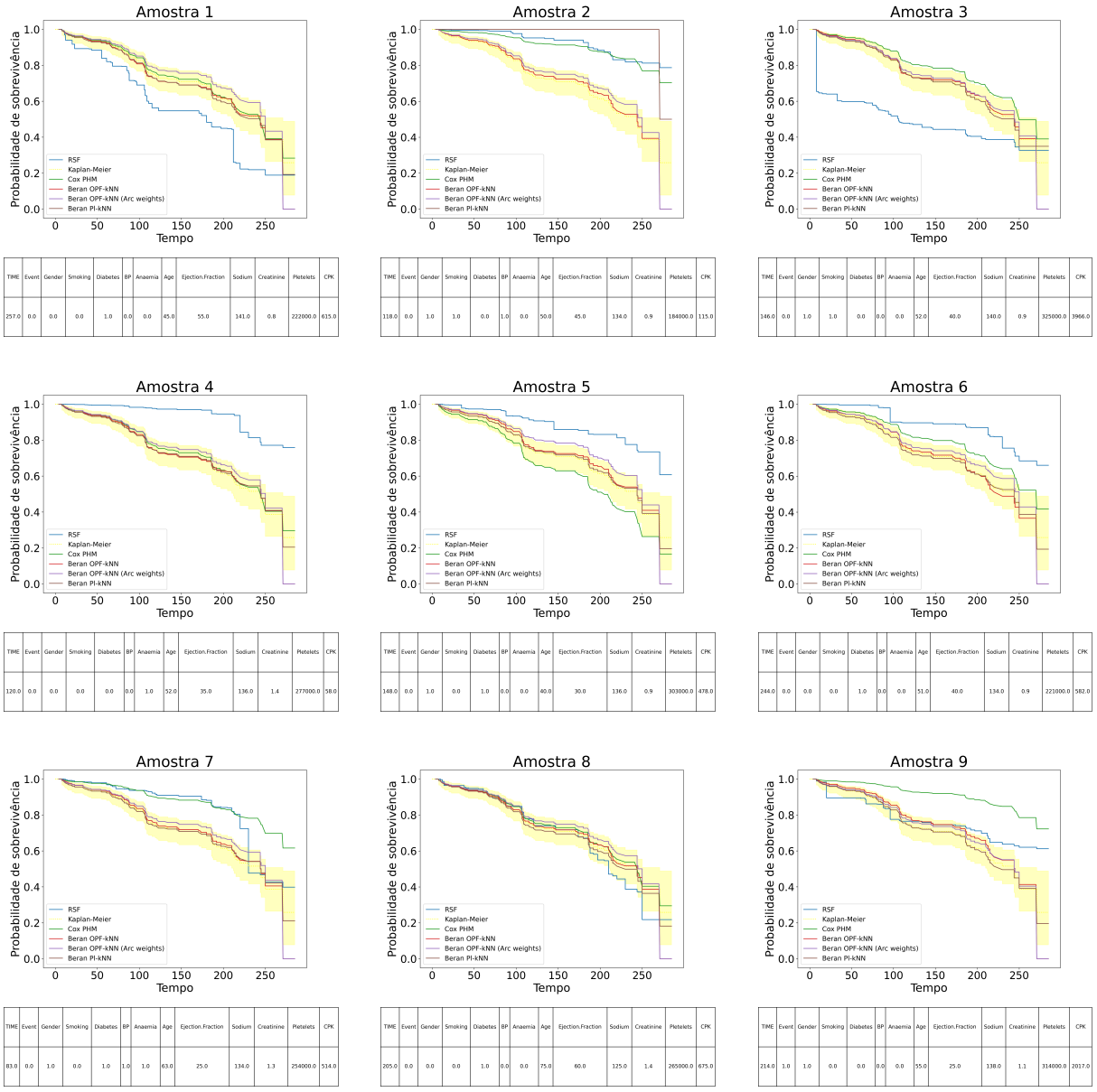
O trio de modelos propostos prova novamente sua resistência à censura através da ausência de tais invariações, mostrando intervalos de constância mais raros e mais curtos que os métodos de Cox e RSF. Entre eles, Beran OPF-kNN e Beran PL-kNN tendem a seguir de forma próxima o Kaplan-Meier, se distanciando ao longo do tempo, ao passo que menos amostras em risco permanecem no conjunto de teste. Enquanto isso, o Beran OPF-kNN (Arc weights) superestima a probabilidade de sobrevivência e todas as combinações exceto para WC médio para fabricante C, ao mesmo tempo que o modelo de Cox subestima a sobrevivência para os fabricantes A e B em níveis baixos e a superestima em níveis altos e para o fabricante C. No contexto deste trabalho, superestimar a probabilidade de sobrevivência significa estimar probabilidades mais altas que decrescem à uma taxa mais lenta do que seria o ideal de acordo com o intervalo do Kaplan-Meier. Em outras palavras, indica uma curva de sobrevivência com um comportamento de decaimento menos acentuado. Por outro lado, subestimar uma probabilidade significa uma curva de sobrevivência com um decaimento mais rápido com relação aos demais modelos.

Todos os modelos de referência começam a sofrer distorções dentro do intervalo entre 4.000 e 6.000 dias conforme as amostras em risco vão diminuindo, porém, deve ser notado que os métodos desenvolvidos apresentam menor distorção em comparação com os demais modelos, mostrando uma resistência a um número baixo de amostras em uma perspectiva qualitativa.

Na figura 3, os novos modelos geram curvas que permanecem dentro do intervalo do estimador de Kaplan-Meier e extremamente similares entre si, produzindo funções com formatos quase idênticos, porém em valores diferentes de probabilidades. Ou seja, as sobrevivências calculadas, de forma geral, variaram quase em mesma quantidade, nos mesmos pontos do tempo, porém partindo de pontos iniciais diferentes. O mesmo fenômeno ocorre com o modelo de Cox nas amostras 1, 3, 4, 5, 6 e 8, mas não nas demais.

O RSF, por sua vez, produz uma predição única, mas com menos distorções que no primeiro conjunto de dados, observando-se um número menor de planícies de constância e menos quedas bruscas ao analisar essas amostras. Além disso, para todas as amostras testadas exceto a amostra 8, ficou bastante distante do Kaplan-Meier, prevendo probabilidades vastamente superiores ou inferiores à do KM em todos os

Figura 3 – Funções de sobrevivência para as nove amostras selecionadas aleatoriamente do conjunto de insuficiência cardíaca.

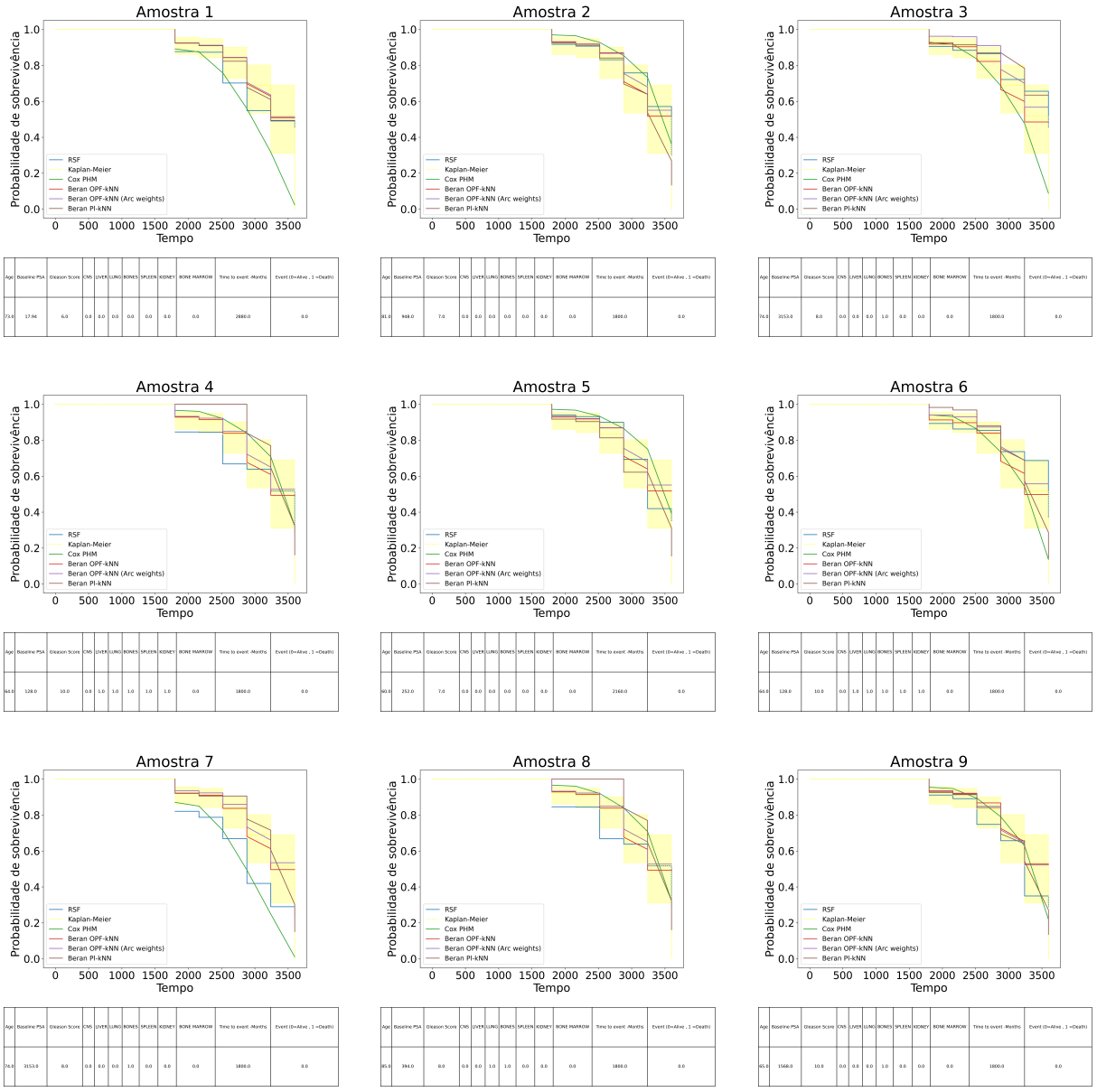


Fonte: Imagens elaboradas pelo autor

outros casos. Porém, a análise quantitativa previamente discutida comprova que há uma maior chance do RSF ter obtido a melhor estimativa, mostrando que a proximidade do Kaplan-Meier nem sempre é sinônimo de uma boa previsão.

Chama a atenção o comportamento do Beran PL-kNN ao analisar a amostra 2, para a qual conferiu uma probabilidade de sobrevivência de 100% até depois de passados 250 dias. Tal fenômeno denuncia uma vulnerabilidade do modelo: um conjunto de dados em que as amostras formam *clusters* muito compactos, possibilitando que a distância até o centróide mais próximo da classe da amostra de teste seja muito

Figura 4 – Funções de sobrevivência das nove amostras selecionadas aleatoriamente do conjunto de câncer de próstata.



Fonte: Imagens elaboradas pelo autor

curta e, consequentemente, o semi-círculo usado pelo PL-kNN para aprender o valor de k seja muito pequeno, prejudicando a previsão para aquela amostra. É razoável supor que o motivo do desempenho quantitativo levemente inferior do modelo em relação aos demais métodos propostos se deva a outros casos como esse.

Agora, na figura 4, os modelos tiveram o desafio de lidar com uma pequena variedade de tempos e o menor número de amostras dos três conjuntos de dados, levando à maiores distorções e a formação de mais ‘degraus’ de probabilidades do que nas outras seleções de amostras e probabilidades que começam a diminuir de

valor apenas entre 1.500 e 2.000 dias. O modelo de Cox se mostrou o mais resistente a esse panorama, mantendo probabilidades dinâmicas em todos os casos, enquanto os modelos desenvolvidos novamente se mantém próximos à estimativa do Kaplan-Meier, com a exceção do Beran PL-kNN nas amostras 3, 4, 6 e 8, nas quais produziu uma curva mais 'otimista', ou seja, uma curva que superestima a sobrevivência, que seus contemporâneos.

No entanto, ainda é possível constatar um grau ligeiramente menor de distorção inclusive ao comparar as curvas dos métodos propostos à gerada pelo RSF, agora muito mais próximas às estimadas pelo KM e composta inteiramente por planícies de probabilidade constante. Sendo assim, os novos modelos exibem novamente a sua capacidade de operar em situações com um número menor de amostras e, portanto, de informação limitada.

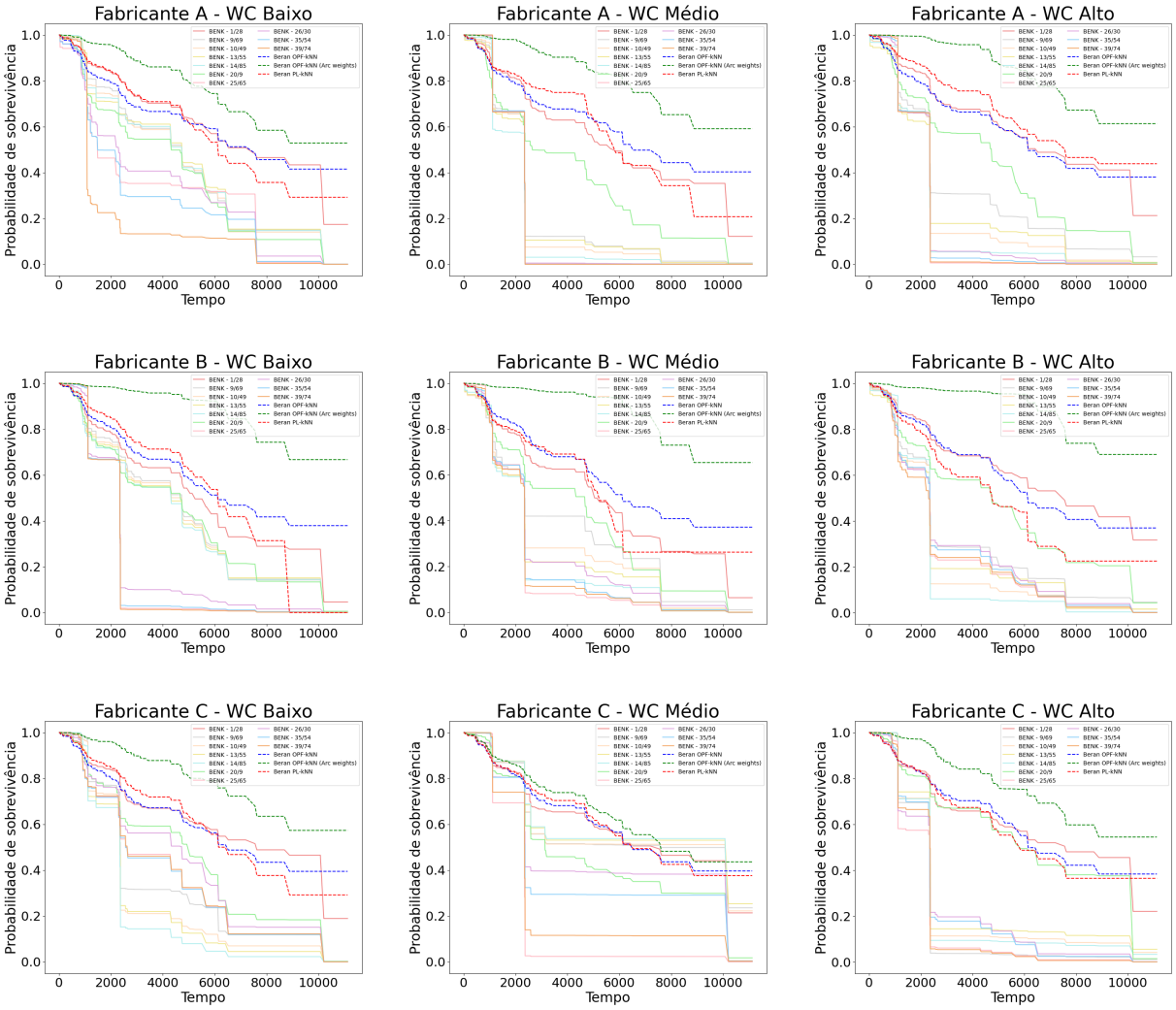
As figuras 5, 6 e 7 mostram as mesmas curvas de sobrevivência obtidas a partir das mesmas amostras pela abordagem proposta, mas em comparação com configurações diferentes do BENK. A grande variedade de resultados mostra a sensibilidade do BENK à seleção de hiperparâmetros, algo que impacta diretamente as probabilidades de sobrevivência ao longo do tempo. Os métodos propostos obtiveram resultados parecidos a várias variações do modelo baseado em *kernels* neurais, ao mesmo tempo que diferiram significantemente de outras.

Baixas quantidades de observações de uma determinada combinação de covariáveis surtem efeito nos resultados do BENK, algo mais facilmente constatado nas amostras de válvulas DHSV, em que combinações com menos representantes, como fabricante C e WC médio ou qualquer amostra com WC alto, geram curvas altamente distorcidas para a grande maioria das versões do estimador. Nesses casos, é comum haver apenas uma configuração entre as testadas que se assemelhe às previsões produzidas pelos modelos propostos, ilustrando a necessidade de rigor na otimização de hiperparâmetros que precede o uso do BENK.

Isso é reforçado ao analisar os gráficos adquiridos a partir dos outros dois conjuntos de dados, para os quais, mesmo em casos nos quais boa parte dos modelos geraram curvas semelhantes, algumas configurações do BENK mostram importantes distorções, com longos intervalos de constância seguidos de quedas bruscas e previsões contraditórias aos demais modelos.

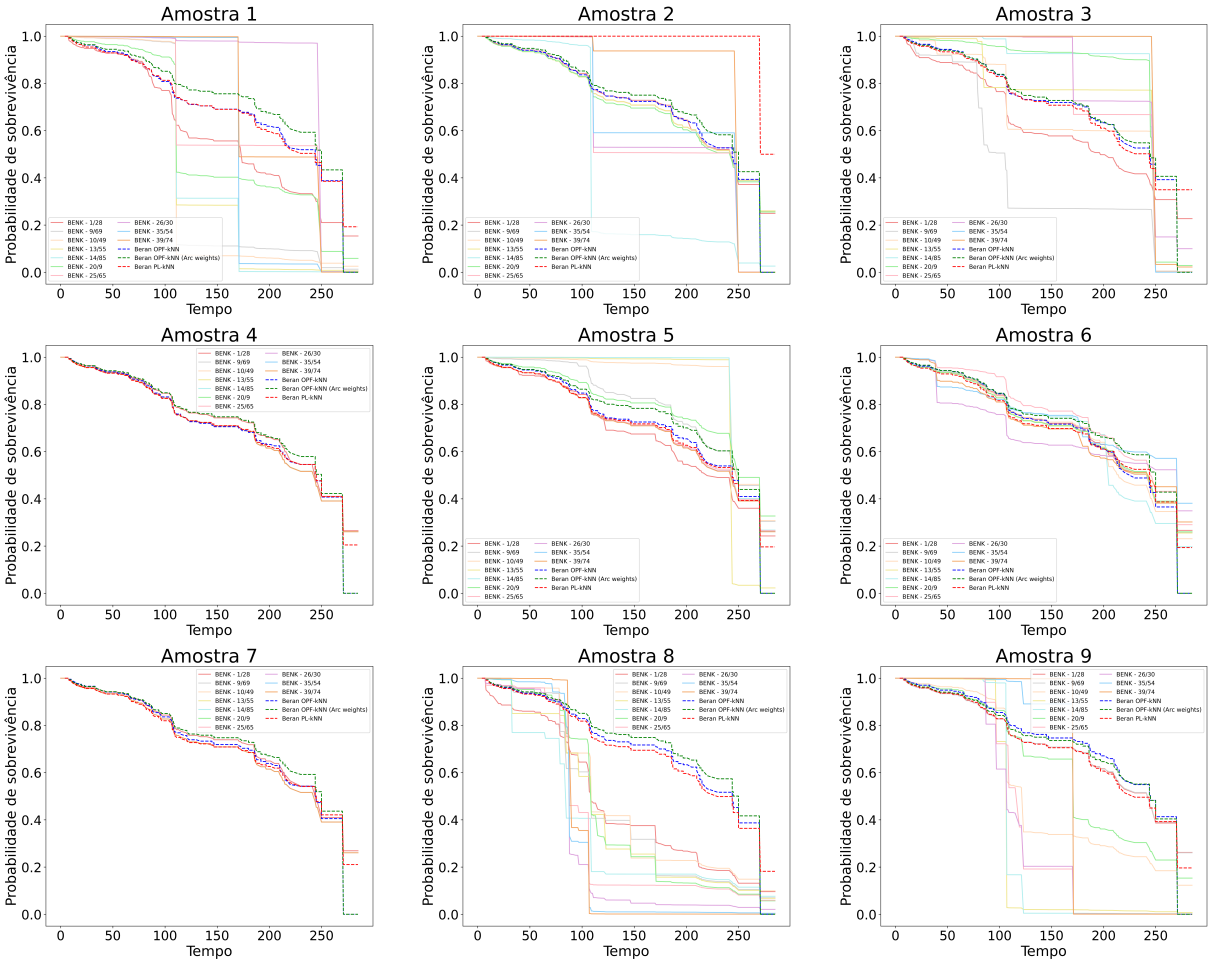
A maioria das funções de sobrevivência estimadas pelas abordagens propostas são semelhantes às do BENK e obtidas com menos requisitos de inicialização de hiperparâmetros. De um ponto de vista qualitativo, esses resultados indicam que os modelos desenvolvidos apresentam menos complexidade ao mesmo tempo que desempenham a tarefa da previsão de sobrevivência com poder discriminativo similar ao modelo BENK.

Figura 5 – Funções de sobrevivência para cada combinação de fabricante e profundidade operacional do conjunto de dados DHSV.



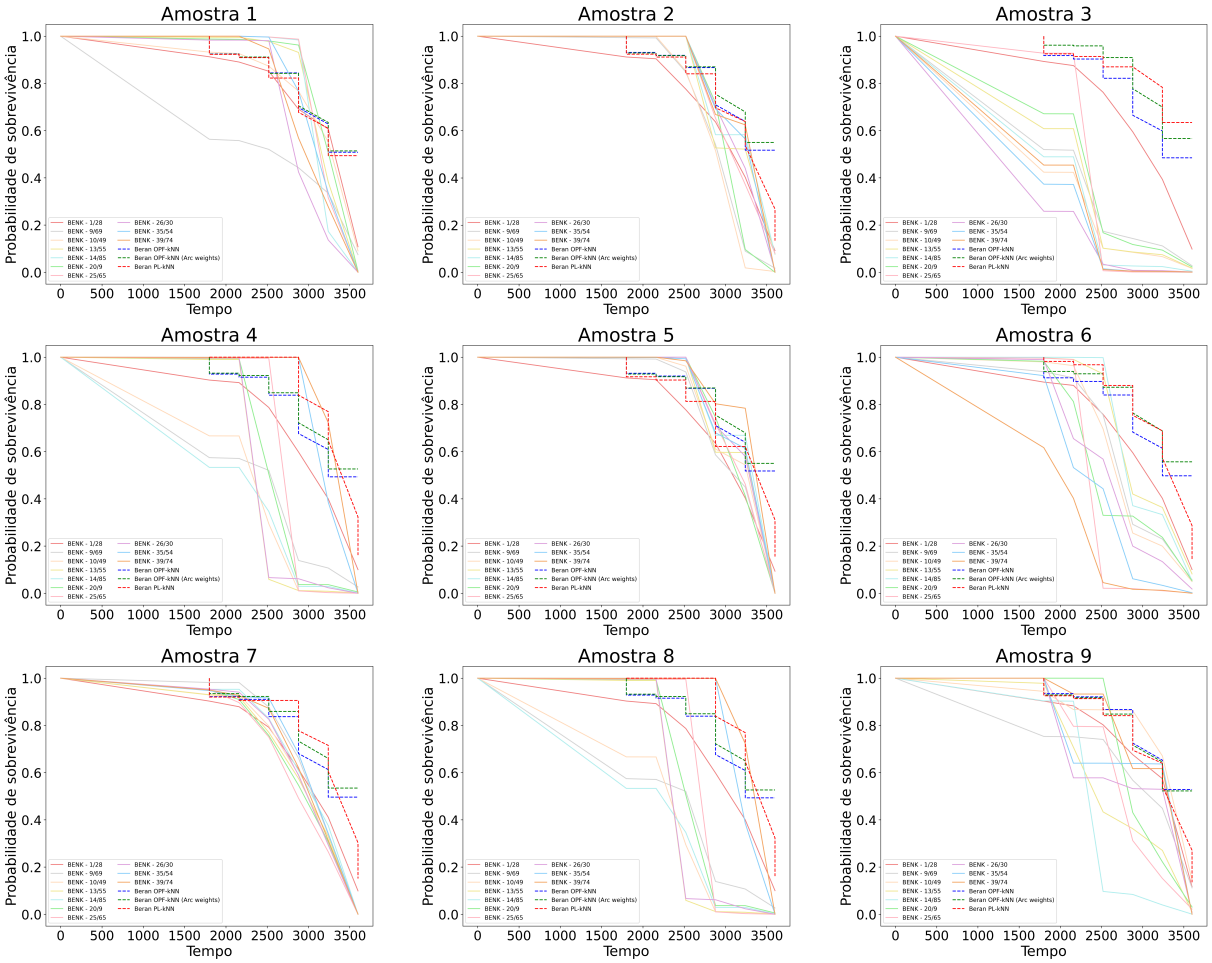
Fonte: Imagens elaboradas pelo autor

Figura 6 – Funções de sobrevivência para as nove amostras selecionadas aleatoriamente do conjunto de insuficiência cardíaca.



Fonte: Imagens elaboradas pelo autor

Figura 7 – Funções de sobrevivência das nove amostras selecionadas aleatoriamente do conjunto de câncer de próstata.



Fonte: Imagens elaboradas pelo autor

5 Conclusão

O presente trabalho apresentou uma abordagem de análise de sobrevivência que combina o estimador de Beran com diversos modelos de aprendizado de máquina ao usá-las como funções de *kernel* para promover a estimativa de sobrevivência não paramétrica e métodos não-paramétricos de aprendizado de máquina. Esse objetivo é alcançado ao adaptar o problema a uma tarefa de classificação internamente ao PL-kNN e ao OPF-kNN, demonstrando o potencial do método proposto e de outras metodologias baseadas em algoritmos de classificação.

Os modelos desenvolvidos expandem a capacidade do estimador de Beran de capturar as relações entre dados complexos e minimizam a otimização de hiperparâmetros, se mostrando como uma forte alternativa a outros métodos baseados em aprendizado de máquina. Ao incorporar de forma robusta os efeitos das estruturas das covariáveis, os modelos refinam a estimativa de probabilidades condicionais de sobrevivência, ao mesmo tempo que descartam o esforço computacional extra de modelos como BENK e o RSF.

No entanto, os modelos demonstraram pontos de vulnerabilidade no que diz respeito a conjuntos de dados com um grande número de covariáveis, formando amostras de alta dimensionalidade que dificultam a execução dos métodos, cujas funções de *kernel* são geradas por algoritmos baseados em distância. Os dados de alta dimensionalidade, entretanto, também apresentavam um número pequeno de amostras, tornando interessante a investigação do desempenho dos algoritmos quando aplicados a conjuntos de alta dimensionalidade e com um grande número de amostras.

Outras possibilidades para trabalhos futuros envolvem o emprego de uma abordagem de regularização de *kernels* que integra as probabilidades de Kaplan-Meier dos vizinhos mais próximos nas funções de distância dos modelos, bem como outras novas funções de *kernel*. Além disso, se mostra interessante uma abordagem multi-*kernel*, em que várias diversas funções de kernel são usadas em um mesmo estimador.

Referências

AHMAD, T.; MUNIR, A.; BHATTI, S. H.; AFTAB, M.; RAZA, M. A. Survival analysis of heart failure patients: A case study. *PloS one*, Public Library of Science San Francisco, CA USA, v. 12, n. 7, p. e0181001, 2017.

ALLOGHANI, M.; AL-JUMEILY, D.; MUSTAFINA, J.; HUSSAIN, A.; ALJAAF, A. J. A systematic review on supervised and unsupervised machine learning algorithms for data science. *Supervised and unsupervised learning for data science*, Springer, p. 3–21, 2020.

ALPAYDIN, E. *Introduction to machine learning*. [S.l.]: MIT press, 2020.

ALSINA, E. F.; CABRI, G.; REGATTIERI, A. A neural network approach to find the cumulative failure distribution: Modeling and experimental evidence. *Quality and reliability engineering international*, Wiley Online Library, v. 32, n. 2, p. 567–579, 2016.

AMERI, S.; FARD, M. J.; CHINNAM, R. B.; REDDY, C. K. Survival analysis based framework for early prediction of student dropouts. In: *Proceedings of the 25th ACM international on conference on information and knowledge management*. [S.l.: s.n.], 2016. p. 903–912.

ARAÚJO, J. M. A. Análise de sobrevivência e previsão de churn de clientes de seguros de vida do banco do brasil. 2022.

BELLE, V. V.; PELCKMANS, K.; HUFFEL, S. V.; SUYKENS, J. A. Support vector methods for survival analysis: a comparison between ranking and regression approaches. *Artificial intelligence in medicine*, Elsevier, v. 53, n. 2, p. 107–118, 2011.

BENDER, A.; RÜGAMER, D.; SCHEIPL, F.; BISCHL, B. A general machine learning framework for survival analysis. In: SPRINGER. *Joint European conference on machine learning and knowledge discovery in databases*. [S.l.], 2020. p. 158–173.

BERAN, R. Nonparametric regression with randomly censored survival data. technical report, University of California, Berkeley, 1981.

BISCHL, B.; BINDER, M.; LANG, M.; PIELOK, T.; RICHTER, J.; COORS, S.; THOMAS, J.; ULLMANN, T.; BECKER, M.; BOULESTEIX, A.-L. et al. Hyperparameter optimization: Foundations, algorithms, best practices, and open challenges. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, Wiley Online Library, v. 13, n. 2, p. e1484, 2023.

BREIMAN, L. Random forests. *Machine learning*, Springer, v. 45, n. 1, p. 5–32, 2001.

CHAKRABORTY, A.; TSOKOS, C. P. Parametric and non-parametric survival analysis of patients with acute myeloid leukemia (aml). *Open Journal of Applied Sciences*, Scientific Research Publishing, v. 11, n. 1, p. 126–148, 2021.

COLOMBO, D. Caracterização do comportamento de confiabilidade das barreiras de segurança de poços de petróleo em diferentes condições de operação utilizando aprendizado de máquina. 2023.

COLOMBO, D.; LIMA, G. B. A.; PAPA, J. P.; PASSOS, L. A.; SANTANA, M. C. S. A novel approach to well barrier survival analysis using machine learning. In: *Proceedings of the 32nd European Safety and Reliability Conference, Dublin, Ireland*. [S.l.: s.n.], 2022. v. 28.

COLOMBO, D.; LIMA, G. B. A.; PEREIRA, D. R.; PAPA, J. P. Regression-based finite element machines for reliability modeling of downhole safety valves. *Reliability Engineering & System Safety*, Elsevier, v. 198, p. 106894, 2020.

COLOSIMO, E. A.; GIOLO, S. R. *Análise de sobrevivência aplicada*. [S.l.]: Editora Blucher, 2024.

CORTES, C.; VAPNIK, V. Support-vector networks. *Machine learning*, Springer, v. 20, n. 3, p. 273–297, 1995.

CROW, E. L.; SHIMIZU, K. *Lognormal distributions*. [S.l.]: Marcel Dekker New York, 1987.

CZADO, C.; RUDOLPH, F. Application of survival analysis methods to long-term care insurance. *Insurance: Mathematics and Economics*, Elsevier, v. 31, n. 3, p. 395–413, 2002.

DIAMOUTENE, A.; BARRO, D.; SOMDA, S. M. A.; NOUREDDINE, F.; KAMSU-FOGUEM, B. Survival analysis in living and engineering sciences. *JP Journal of Biostatistics*, v. 13, n. 2, p. 223–238, 2016.

FALCAO, A. X.; STOLFI, J.; LOTUFO, R.; CUNHA, B. da; ARAUJO, G. Image foresting transform. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1999.

FALKNER, S.; KLEIN, A.; HUTTER, F. Bohb: Robust and efficient hyperparameter optimization at scale. In: PMLR. *International conference on machine learning*. [S.l.], 2018. p. 1437–1446.

FARD, M. J.; WANG, P.; CHAWLA, S.; REDDY, C. K. A bayesian perspective on early stage event prediction in longitudinal data. *IEEE Transactions on Knowledge and Data Engineering*, IEEE, v. 28, n. 12, p. 3126–3139, 2016.

GRAF, E.; SCHMOOR, C.; SAUERBREI, W.; SCHUMACHER, M. Assessment and comparison of prognostic classification schemes for survival data. *Statistics in medicine*, Wiley Online Library, v. 18, n. 17-18, p. 2529–2545, 1999.

GUPTA, R. C.; KANNAN, N.; RAYCHAUDHURI, A. Analysis of lognormal survival data. *Mathematical biosciences*, Elsevier, v. 139, n. 2, p. 103–115, 1997.

HARRELL JR, F. E.; HARRELL, F. E. Cox proportional hazards regression model. *Regression modeling strategies: With applications to linear models, logistic and ordinal regression, and survival analysis*, Springer, p. 475–519, 2015.

HASTIE, T. *The elements of statistical learning: data mining, inference, and prediction*. [S.l.]: Springer, 2009.

HUANG, Y.; LI, J.; LI, M.; APARASU, R. R. Application of machine learning in predicting survival outcomes involving real-world data: a scoping review. *BMC medical research methodology*, Springer, v. 23, n. 1, p. 268, 2023.

ISHWARAN, H.; KOGALUR, U. B.; BLACKSTONE, E. H.; LAUER, M. S. Random survival forests. 2008.

IZBICKI, R.; SANTOS, T. M. dos. *Aprendizado de máquina: uma abordagem estatística*. [S.l.]: Rafael Izbicki, 2020.

JAMES, G.; WITTEN, D.; HASTIE, T.; TIBSHIRANI, R. *An introduction to statistical learning: with applications in R*. [S.l.]: Springer, 2013. v. 103.

JODAS, D. S.; PASSOS, L. A.; ADEEL, A.; PAPA, J. P. Pl-k nn: A parameterless nearest neighbors classifier. In: IEEE. *2022 29th International Conference on Systems, Signals and Image Processing (IWSSIP)*. [S.l.], 2022. p. 1–4.

JODAS, D. S.; PASSOS, L. A.; ADEEL, A.; PAPA, J. P. Pl-knn: A python-based implementation of a parameterless k-nearest neighbors classifier. *Software impacts*, Elsevier, v. 15, p. 100459, 2023.

JR, F. E. H.; LEE, K. L.; MARK, D. B. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in medicine*, Wiley Online Library, v. 15, n. 4, p. 361–387, 1996.

KAPLAN, E. L.; MEIER, P. Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, Taylor & Francis, v. 53, n. 282, p. 457–481, 1958.

KATONGOLE, P.; SANDE, O. J.; NIYONZIMA, N. Dataset, *Survival analysis data for metastatic prostate cancer patients at UCI*. figshare, 2021. Disponível em: <https://doi.org/10.6084/m9.figshare.13625627.v1>. Acesso em: 17/11/2025.

KHAN, F. M.; ZUBEK, V. B. Support vector regression for censored data (svrc): a novel tool for survival analysis. In: IEEE. *2008 Eighth IEEE International Conference on Data Mining*. [S.l.], 2008. p. 863–868.

KIESSLING, J.; BRUNNBERG, A.; HOLTE, G.; ELDRUP, N.; SÖRELIUS, K. Artificial intelligence outperforms kaplan–meier analyses estimating survival after elective treatment of abdominal aortic aneurysms. *European Journal of Vascular and Endovascular Surgery*, v. 65, n. 4, p. 600–607, 2023.

KIM, D. W.; LEE, S.; KWON, S.; NAM, W.; CHA, I.-H.; KIM, H. J. Deep learning-based survival prediction of oral cancer patients. *Scientific reports*, Nature Publishing Group UK London, v. 9, n. 1, p. 6994, 2019.

KIRPICHENKO, S.; UTKIN, L.; KONSTANTINOV, A.; MULIUKHA, V. Benk: The beran estimator with neural kernels for estimating the heterogeneous treatment effect. *Algorithms*, MDPI, v. 17, n. 1, p. 40, 2024.

KUHN, M.; JOHNSON, K. et al. *Applied predictive modeling*. [S.l.]: Springer, 2013. v. 26.

LAI, C.-D.; MURTHY, D.; XIE, M. Weibull distributions and their applications. In: *Springer handbook of engineering statistics*. [S.l.]: Springer, 2006. p. 63–78.

LAWLESS, J. F. *Statistical models and methods for lifetime data*. [S.l.]: John Wiley & Sons, 2011.

LEE, E. T.; GO, O. T. Survival analysis in public health research. *Annual review of public health*, Annual Reviews 4139 El Camino Way, PO Box 10139, Palo Alto, CA 94303-0139, USA, v. 18, n. 1, p. 105–134, 1997.

LEE, E. T.; WANG, J. *Statistical methods for survival data analysis*. [S.l.]: John Wiley & Sons, 2003. v. 476.

LEE, W.-M. *Python machine learning*. [S.l.]: John Wiley & Sons, 2019.

LI, L.; JAMIESON, K.; DESALVO, G.; ROSTAMIZADEH, A.; TALWALKAR, A. Hyperband: Bandit-based configuration evaluation for hyperparameter optimization. In: *International Conference on Learning Representations*. [S.l.: s.n.], 2022.

LIMONGI, J. E.; SANTOS, K. A. R.; PERISSATO, I. L.; PINTO, R. d. M. C.; MENDONÇA, T. M. d. S.; RINALDI, A. E. M. Survival analysis of chagas disease patients, beneficiaries of social security and social assistance in brazil, 1942–2016. *Revista Brasileira de Epidemiologia*, SciELO Public Health, v. 27, p. e240020, 2024.

LÓPEZ-CHEDA, A.; CAO, R.; JÁCOME, M. A.; KEILEGOM, I. V. Nonparametric incidence estimation and bootstrap bandwidth selection in mixture cure models. *Computational Statistics & Data Analysis*, Elsevier, v. 105, p. 144–165, 2017.

LOUZADA, F.; CUMINATO, J. A.; RODRIGUEZ, O. M.; TOMAZELLA, V. L.; FERREIRA, P. H.; RAMOS, P. L.; NIAKI, S. R.; GONZATTO, O. A.; PERISSINI, I. C.; ALEGRIA, L. F. et al. A repairable system subjected to hierarchical competing risks: Modeling and applications. *IEEE Access*, IEEE, v. 7, p. 171707–171723, 2019.

LOUZADA, F.; CUMINATO, J. A.; RODRIGUEZ, O. M. H.; TOMAZELLA, V. L.; MILANI, E. A.; FERREIRA, P. H.; RAMOS, P. L.; BOCHIO, G.; PERISSINI, I. C.; JUNIOR, O. A. G. et al. Incorporation of frailties into a non-proportional hazard regression model and its diagnostics for reliability modeling of downhole safety valves. *IEEE Access*, IEEE, v. 8, p. 219757–219774, 2020.

MARSHALL, A. W.; OLKIN, I. A multivariate exponential distribution. *Journal of the American Statistical Association*, Taylor & Francis, v. 62, n. 317, p. 30–44, 1967.

MEEKER, W. Q.; ESCOBAR, L. A.; PASCUAL, F. G. *Statistical methods for reliability data*. [S.l.]: John Wiley & Sons, 2021.

MODARRES, M.; KAMINSKIY, M. P.; KRIVTSOV, V. *Reliability engineering and risk analysis: a practical guide*. [S.l.]: CRC press, 2016.

MURTHY, D. P.; XIE, M.; JIANG, R. *Weibull models*. [S.l.]: John Wiley & Sons, 2004.

NEWAZ, A.; HAQ, F. S.; AHMED, N. A case study on risk prediction in heart failure patients using random survival forest. In: *2021 5th International Conference on Electrical Engineering and Information & Communication Technology (ICEEICT)*. [S.l.]: IEEE, 2021. p. 1–6.

PAPA, J. P.; FALCAO, A. X.; SUZUKI, C. T. Supervised pattern classification based on optimum-path forest. *International Journal of Imaging Systems and Technology*, Wiley Online Library, v. 19, n. 2, p. 120–131, 2009.

PAPATHANASIOU, D.; DEMERTZIS, K.; TZIRITAS, N. Machine failure prediction using survival analysis. *Future Internet*, MDPI, v. 15, n. 5, p. 153, 2023.

PRINCE, T.; BOMMERT, A.; RAHNENFÜHRER, J.; SCHMID, M. On the estimation of inverse-probability-of-censoring weights for the evaluation of survival prediction error. *PloS one*, Public Library of Science San Francisco, CA USA, v. 20, n. 1, p. e0318349, 2025.

RASHEED, M. Analyzing applications and properties of the exponential continuous distribution in reliability and survival analysis. *Journal of Positive Sciences*, v. 4, n. 5, p. 71–79, 2023.

SANTOS, H. G. d. *Comparação da performance de algoritmos de machine learning para a análise preditiva em saúde pública e medicina*. Tese (Doutorado) — Universidade de São Paulo, 2018.

SCHOBER, P.; VETTER, T. R. Kaplan-meier curves, log-rank tests, and cox regression for time-to-event data. *Anesthesia & Analgesia*, LWW, v. 132, n. 4, p. 969–970, 2021.

SHAMSUTDINOVA, D.; STAMATE, D.; ROBERTS, A.; STAHL, D. Combining cox model and tree-based algorithms to boost performance and preserve interpretability for health outcomes. In: SPRINGER. *IFIP International Conference on Artificial Intelligence Applications and Innovations*. [S.l.], 2022. p. 170–181.

SHOURABIZADEH, H.; ALEMAN, D. M.; ROUSSEAU, L.-M.; ZHENG, K.; BHAT, M. Classification-augmented survival estimation (case): A novel method for individualized long-term survival prediction with application to liver transplantation. *PLOS ONE*, v. 20, n. 1, p. e0315928, 2025.

SNIDER, B.; MCBEAN, E. A. Combining machine learning and survival statistics to predict remaining service life of watermains. *Journal of infrastructure systems*, American Society of Civil Engineers, v. 27, n. 3, p. 04021019, 2021.

SNOEK, J.; LAROCHELLE, H.; ADAMS, R. P. Practical bayesian optimization of machine learning algorithms. *Advances in neural information processing systems*, v. 25, 2012.

SPOONER, A.; CHEN, E.; SOWMYA, A.; SACHDEV, P.; KOCHAN, N. A.; TROLLOR, J.; BRODATY, H. A comparison of machine learning methods for survival analysis of high-dimensional clinical data for dementia prediction. *Scientific reports*, Nature Publishing Group UK London, v. 10, n. 1, p. 20410, 2020.

THORSEN-MEYER, H.-C.; PLACIDO, D.; KAAS-HANSEN, B. S.; NIELSEN, A. P.; LANGE, T.; NIELSEN, A. B.; TOFT, P.; SCHIERBECK, J.; STRØM, T.; CHMURA, P. J.; HEIMANN, M.; BELLING, K.; PERNER, A.; BRUNAK, S. Discrete-time survival analysis in the critically ill: a deep learning approach using heterogeneous data. *npj Digital Medicine*, v. 5, n. 1, p. 142, 2022.

UNO, H.; CAI, T.; PENCINA, M. J.; D'AGOSTINO, R. B.; WEI, L.-J. On the c-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Statistics in medicine*, Wiley Online Library, v. 30, n. 10, p. 1105–1117, 2011.

VAJARGAH, K. F. Evaluating the parametric, semi-parametric and nonparametric models for reliability in aircraft braking system. *Mathematical Sciences*, Springer, v. 15, n. 3, p. 259–267, 2021.

WANG, P.; LI, Y.; REDDY, C. K. Machine learning for survival analysis: A survey. *ACM Computing Surveys (CSUR)*, ACM New York, NY, USA, v. 51, n. 6, p. 1–36, 2019.

ZHONG, C.; TIBSHIRANI, R. Survival analysis as a classification problem. *arXiv preprint arXiv:1909.11171*, 2019.

ZHU, X.; YAO, J.; HUANG, J. Deep convolutional neural network for survival analysis with pathological images. In: IEEE. *2016 IEEE international conference on bioinformatics and biomedicine (BIBM)*. [S.l.], 2016. p. 544–547.