UNIVERSIDADE FEDERAL DO CEARÁ

CENTRO DE CIÊNCIAS

DEPARTAMENTO DE ESTATÍSTICA E MATEMÁTICA APLICADA

PROGRAMA DE PÓS-GRADUAÇÃO EM MODELAGEM E MÉTODOS QUANTITATIVOS

MESTRADO ACADÊMICO EM MODELAGEM E MÉTODOS QUANTITATIVOS

ISABEL DE CASTRO BENEYTO

UMA ABORDAGEM BASEADA EM SIMILARIDADE EMPÍRICA PARA O ESTIMADOR DE KAPLAN-MEIER

FORTALEZA

2024

ISABEL DE CASTRO BENEYTO

UMA ABORDAGEM BASEADA EM SIMILARIDADE EMPÍRICA PARA O ESTIMADOR DE KAPLAN-MEIER

Dissertação apresentada ao Curso de Mestrado Acadêmico em Modelagem e Métodos Quantitativos do Programa de Pós-Graduação em Modelagem e Métodos Quantitativos da Universidade Federal do Ceará, como requisito parcial à obtenção do título de mestra em Modelagem e Métodos Quantitativos. Área de Concentração: Modelagem e Métodos Quantitativos.

Orientador: Prof. Dr. Leandro Chaves Rêgo.
Coorientador: Prof. Dr. Anselmo R. Pitombeira Neto.

FORTALEZA

2024

Dados Internacionais de Catalogação na Publicação
Universidade Federal do Ceará
Sistema de Bibliotecas
Gerada automaticamente pelo módulo Catalog, mediante os dados fornecidos pelo(a) autor(a)

B398a Beneyto, Isabel de Castro.
Uma abordagem baseada em similaridade empírica para o estimador de Kaplan-Meier / Isabel de Castro Beneyto. – 2024.
87 f. : il. color.

Dissertação (mestrado) – Universidade Federal do Ceará, Centro de Ciências, Programa de Pós-Graduação em Modelagem e Métodos Quantitativos, Fortaleza, 2024.
Orientação: Prof. Dr. Leandro Chaves Rêgo.
Coorientação: Prof. Dr. Anselmo R. Pitombeira Neto.

1. Análise de sobrevivência (biometria). 2. Teoria da estimativa. 3. Similaridade empírica. I. Título.
CDD 510

ISABEL DE CASTRO BENEYTO

UMA ABORDAGEM BASEADA EM SIMILARIDADE EMPÍRICA PARA O ESTIMADOR DE KAPLAN-MEIER

Dissertação apresentada ao Curso de Mestrado Acadêmico em Modelagem e Métodos Quantitativos do Programa de Pós-Graduação em Modelagem e Métodos Quantitativos da Universidade Federal do Ceará, como requisito parcial à obtenção do título de mestra em Modelagem e Métodos Quantitativos. Área de Concentração: Modelagem e Métodos Quantitativos.

Aprovada em: 22/05/2024.

BANCA EXAMINADORA

Prof. Dr. Leandro Chaves Rêgo (Orientador)
Universidade Federal do Ceará (UFC)

Prof. Dr. Anselmo R. Pitombeira Neto (Coorientador)
Universidade Federal do Ceará (UFC)

Prof. Dr. Juvêncio Santos Nobre
Universidade Federal do Ceará (UFC)

Prof. Dr. Raydonal Ospina Martinez
Universidade Federal da Bahia (UFBA)

Aos meus pais.

AGRADECIMENTOS

Em primeiro lugar, expresso minha gratidão à minha família, especialmente à minha mãe, Francisca Leite de Castro Beneyto, e ao meu pai, Juan Tomas Beneyto Paysal, pelo constante apoio, suporte e incentivo. Além disso, agradeço aos meus cachorros Bolt e Jack por todo o amor incondicional, alegria e companheirismo que eles trazem à minha vida diariamente. É verdadeiramente um privilégio contar com uma família tão amorosa.

Ao Prof. Dr. Leandro Chaves Rêgo, agradeço pela orientação excepcional, paciência e dedicação ao meu trabalho. Ao meu coorientador, Dr. Anselmo R. Pitombeira Neto, agradeço por todos os ensinamentos valiosos e pelo comprometimento durante a pesquisa. Também expresso minha gratidão a todos os membros do Departamento de Estatística e Matemática Aplicada, incluindo funcionários e professores, pela rica contribuição para o meu conhecimento e formação.

Aos professores que participaram da banca examinadora, Dr. Juvêncio Santos Nobre e Dr. Raydonal Ospina Martinez, agradeço pelo tempo dedicado, pelas colaborações valiosas e pelas sugestões recebidas.

Por último, mas não menos importante, agradeço à FUNCAP pelo apoio financeiro que possibilitou a manutenção da bolsa de auxílio e pelo estímulo à pesquisa científica ao longo do meu mestrado.

“Sucesso é ir de fracasso em fracasso sem perder o entusiasmo” (Winston Churchill).

RESUMO

O estimador de Kaplan-Meier é amplamente utilizado para estimar a curva de sobrevivência de uma população, incorporando a possibilidade de existência de dados censurados. Embora seja comum em estudos de sobrevida, especialmente em pesquisas clínicas e epidemiológicas, o estimador não leva diretamente em conta as covariáveis em suas previsões. Neste trabalho, propomos uma adaptação do estimador de Kaplan-Meier, denominada estimador de Kaplan-Meier baseado em similaridade. Essa adaptação inclui um quantificador de similaridade na fórmula padrão do estimador, permitindo a atribuição de pesos às covariáveis. Esses pesos são estimados a partir dos dados usando uma função de similaridade predefinida e são obtidos por meio da máxima verossimilhança empírica. Demonstramos a aplicação desse método para prever curvas de sobrevivência condicionais por meio de simulações em diferentes configurações e cenários. As análises foram feitas avaliando os pesos estimados, a variabilidade dos tempos de falha estimados através do desvio padrão e o desempenho do estimador em duas métricas de avaliação: índice de concordância (CI) e Brier Score (BS). Realizamos simulações usando duas formas para a função de similaridade: exponencial (EX) e fracionária (FR). Avaliamos o efeito da normalização dos pesos estimados, a escolha entre o tempo médio e mediano para estimar o tempo de falha a partir da curva de sobrevivência e diferentes medidas de distância para compor a função de similaridade. Para avaliar o desempenho do estimador em diferentes contextos, estimamos os pesos em diferentes amostras de dados, variando o tamanho e a taxa de censura. Finalmente, também realizamos comparações diretas entre o desempenho do estimador proposto e diversos métodos estatísticos e de aprendizado de máquina que são referência no contexto de análise de sobrevivência. Os resultados indicam que o estimador apresenta desempenho competitivo e consistente. Em comparação com os métodos estatísticos, o estimador proposto destaca-se, pois tem a vantagem de não assumir distribuição específica ou riscos proporcionais. Por outro lado, comparado aos algoritmos de aprendizado de máquina, alcançamos uma interpretação mais simples dos parâmetros estimados e evitamos os problemas de sobreajuste dos dados frequentemente associados a modelos excessivamente complexos.

Palavras-chave: análise de sobrevivência (biometria); teoria da estimativa; similaridade empírica.

ABSTRACT

The Kaplan-Meier estimator is widely used for estimating the survival curve of a population, incorporating the possibility of censored data. Although common in survival studies, especially in clinical and epidemiological research, the estimator does not directly consider covariates in its predictions. In this work, we propose adapting the Kaplan-Meier estimator, called the similarity-based Kaplan-Meier estimator. This adaptation includes a similarity quantifier in the standard formula of the estimator, allowing for the assignment of weights to covariates. These weights are estimated from the data using a predefined similarity function and are obtained through empirical maximum likelihood. We demonstrate the application of this method to predict conditional survival curves through simulations in different settings and scenarios. Analyses were performed evaluating the estimated weights, the variability of estimated failure times through standard deviation, and the estimator's performance on two evaluation metrics: concordance index (CI) and Brier Score (BS). We conducted simulations using two forms for the similarity function: exponential (EX) and fractional (FR). We evaluated the effect of normalizing the estimated weights, the choice between mean and median time to estimate failure time from the survival curve, and different distance measures to compose the similarity function. To assess the performance of the estimator in different contexts, we estimated the weights in different data samples, varying the size and censoring rate. Finally, we also performed direct comparisons between the performance of the proposed estimator and various statistical and machine learning methods that are referenced in the context of survival analysis. The results indicate that the estimator demonstrates competitive and consistent performance. Compared to statistical methods, the proposed estimator stands out because it does not assume a specific distribution or proportional hazards. On the other hand, compared to machine learning algorithms, we achieve a simpler interpretation of the estimated parameters and avoid the overfitting issues often associated with overly complex models.

Keywords: survival analysis (biometry); estimation theory; empirical similarity.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

BS	Brier Score
CI	Concordance Index
DBP	Distância Binária Ponderada
DEN	Distância Euclidiana Normalizada
DEP	Distância Euclidiana Ponderada
DMP	Distância de Minkowski Ponderada
IBS	Integrated Brier Score
KM	Kaplan-Meier
SBKM	Similarity-based Kaplan-Meier

LISTA DE SÍMBOLOS

t	Tempo
t_m	Tempo médio
$t_{0.5}$	Tempo mediano
$S(t)$	Função de sobrevivência
$F(t)$	Função de distribuição acumulada
$h(t)$	Função de risco
$\mathbf{w}$	Vetor de pesos associados às covariáveis
$\mathbf{x}$	Vetor de covariáveis
s_w	Função de similaridade
$L(w)$	Função de verossimilhança

SUMÁRIO

1 INTRODUÇÃO

Na pesquisa médica, é comum a existência de dados que representam os tempos de vida de pacientes submetidos a tratamentos médicos ou diagnosticados com alguma doença. No século XIX, tabelas de vida eram usadas em dados provenientes de censos demográficos para investigar padrões de mortalidade e estimar a expectativa de vida dos seres humanos (Kasumigaseki, 2005). Dados desta natureza são chamados de "dados de sobrevivência", pois registram o tempo até a ocorrência de um evento de interesse qualquer.

Originalmente, a análise de sobrevivência preocupava-se em prever quando um paciente viria a óbito, assim como identificar os fatores que influenciariam no tempo de sobrevivência. Entretanto, hoje os métodos são aplicados em diversas áreas. Na Engenharia, onde se utiliza o termo "análise de confiabilidade", é possível avaliar o risco de falha de componentes mecânicos e elétricos (Yang *et al.*, 2022). Em Sociologia, os métodos são utilizados para prever a taxa de desemprego, retenção acadêmica e outras questões sociais (Box-Steffensmeier *et al.*, 2015). Por exemplo, criminalistas estudam o tempo entre a liberação de presos e a ocorrência de novos crimes, estudando os riscos associados a reincidência (Benda, 2003). Portanto, existe uma ampla variedade de campos para estudar e analisar dados de tempo, como também realizar previsões sobre eventos futuros.

Contudo, quando o resultado de um estudo é o tempo até a ocorrência de um evento, existem algumas limitações que precisam ser consideradas. Um dos desafios técnicos em análise de sobrevivência é que, normalmente, ao coletar dados, nem sempre é possível observar o evento de interesse para cada unidade amostral. Por exemplo, pode não ser possível esperar até que todos os pacientes em um estudo clínico tenham morrido ou que recebam alta. Alguns pacientes podem ter inclusive saído do estudo antes do tempo – foram perdidos na pesquisa. Logo, a única informação que temos sobre alguns indivíduos é que eles ainda estavam vivos no último acompanhamento.

Em nossos dados, frequentemente vamos observar a duração do tempo que pretendemos estudar apenas para algumas unidades amostrais, mas não para todas. Quando não conseguimos observar a duração completa para uma unidade amostral, dizemos que essa unidade possui uma observação censurada. Isto difere dos problemas tradicionais de classificação e regressão, onde a variável que queremos prever é conhecida para todas as instâncias de dados.

Existem diferentes abordagens para analisar dados censurados, incluindo métodos não paramétricos, métodos semi-paramétricos e métodos paramétricos. Além disso, muitos

algoritmos de aprendizado de máquina também foram adaptados para lidar com a censura (Wang *et al.*, 2019). A escolha do método depende, principalmente, do contexto dos dados e da pesquisa que está sendo tratada.

Dentre as técnicas, o estimador de Kaplan-Meier, também conhecido como o estimador de limite de produto, é um procedimento não paramétrico utilizado para estimar a função de sobrevivência de uma população, que é a probabilidade de um evento de interesse (como morte, falha ou ocorrência de uma doença) não ter ocorrido até um determinado instante de tempo. O estimador é baseado nos tempos de sobrevivência observados de uma determinada amostra e pode ser usado para estimar e comparar os tempos de sobrevivência de diferentes grupos, como indivíduos recebendo diferentes tratamentos ou com diferentes fatores de risco. Por ser um procedimento não paramétrico, o modelo é flexível e não assume que os dados seguem uma distribuição de probabilidades particular (Colosimo; Giolo, 2006).

Todavia, uma das limitações do estimador KM é que ele não incorpora covariáveis em suas previsões. Isso implica que a relação entre a variável dependente e as covariáveis não é modelada ou considerada no processo de previsão. Em outras palavras, o estimador de Kaplan-Meier faz previsões com base apenas na variável dependente, sem levar em conta possíveis influências de outras variáveis relevantes. Uma alternativa para contornar essa restrição é estratificar os dados por uma ou mais covariáveis e então obter as estimativas das funções de sobrevivência separadamente para cada estrato. Dessa forma, é possível utilizar testes estatísticos, como o teste de Log-Rank, para comparar as funções de sobrevivência entre os grupos (Mehrotra; West, 2020).

Por exemplo, suponha um conjunto de dados de tempos de sobrevivência e uma variável categórica chamada “tratamento”, que assume os valores “A” ou “B”. Neste caso, é possível estratificar os dados pela variável de tratamento e calcular a estimativa de Kaplan-Meier separadamente para cada estrato. Isso permitiria comparar a função de sobrevivência entre os dois grupos de tratamento. Porém, se temos covariáveis quantitativas ou covariáveis qualitativas com muitas categorias, esse processo se torna inviável.

Na literatura existem algumas propostas para o refinamento do estimador de Kaplan-Meier. Beran (1981) adota uma abordagem não paramétrica usando vizinhos mais próximos, em inglês *k-Nearest Neighbor* (k-NN), e kernels. A ideia constitui-se em incorporar diretamente as covariáveis ao estimador de Kaplan-Meier . Para um indivíduo com um vetor de covariáveis que se deseja fazer uma previsão, devemos primeiro encontrar indivíduos conhecidos cujas

covariáveis são suficientemente semelhantes (por exemplo, escolha o k mais próximo). O estimador de Kaplan-Meier então é aplicado somente a esses sujeitos próximos para estimar a função de probabilidade de sobrevivência condicional. Neste artigo, são fornecidos resultados de consistência para essas estimativas de k-NN, denominado *k-NN survival estimator*, e kernel, denominado *Kernel survival estimator*, para a função de sobrevivência. Os primeiros limites de erro não assintóticos foram estabelecidos para estes estimadores, assim como análises numéricas de desempenho em diferentes conjunto de dados por Chen (2019).

Por outro lado, Hu; Huffer (2019) discutem a dificuldade de lidar com estudos de saúde pública, em que geralmente são coletados dados de diferentes locais e o estimador tradicional simplesmente estima as curvas de sobrevivência marginais usando estratificação. Neste artigo, propõe-se o uso de uma regressão ponderada geograficamente para adicionar pesos geográficos às observações e assim obter versões modificadas do estimador de Kaplan–Meier para representar a curva de sobrevivência do local.

Baseando-se nas abordagens propostas por Beran (1981), Chen (2024) estima as funções de sobrevivência individuais com a ajuda de uma função de kernel capaz de medir quão similar são duas instâncias de dados. Essa função é estimada usando modelos de *deep learning* adaptados à análise de sobrevivência, chamados *Deep Kernel Survival Models*. No artigo, é apresentado um modelo escalável de *deep learning* baseado em kernel chamado *Survival Kernet*, dimensionado para grandes conjuntos de dados de uma maneira que seja passível de interpretação e também de análise teórica. Com este propósito, é mencionada uma adaptação ao estimador de Kaplan-Meier, chamada *conditional Kaplan-Meier estimator*, que é um preditor de kernel.

Por outra perspectiva, Bladt; Furrer (2024) tratam dados mais complexos que contêm, além de censuras, várias formas de contaminação. Por exemplo, em seguros de saúde e invalidez algumas reivindicações de seguro ainda não foram liquidadas, enquanto outras foram falsamente liquidadas e, portanto, serão provavelmente retomadas posteriormente. Neste artigo, os autores estudam como lidar com a contaminação dos dados por meio da integração de conhecimento especializado à análise de sobrevivência não paramétrica clássica, isto é, ao estimador KM. Propõe-se, então, o estimador especializado de Kaplan-Meier que pode ser aplicado a qualquer cenário de análise de sobrevivência para o qual os dados totalmente observados podem ser subestimados e um especialista tem conhecimento sobre quais pontos de dados têm esse recurso ou em que medida.

Neste trabalho, estamos contribuindo para o tema ao apresentar uma nova modi-

ficação do estimador de Kaplan-Meier, incorporando um quantificador de similaridade. Um quantificador de similaridade é geralmente um tipo de pontuação que atribui um valor numérico a um par de sequências com base em sua proximidade. As medidas de similaridade desempenham um papel importante em problemas de previsão, com aplicações abrangentes em aprendizado estatístico, mineração de dados, bioestatística, finanças e diversas outras áreas.

Assim, utilizaremos uma metodologia de previsão e interpretação diferente, mas que se assemelha aos estimadores baseados em kernel mencionados anteriormente. O conceito de similaridade empírica sugere um método estatístico baseado em raciocínio por analogias, em que a previsão de eventos futuros é feita com base em casos anteriores, ou seja, experiências passadas, em oposição ao raciocínio baseado em regras (Gayer *et al.*, 2007). Então, se estamos tentando prever o valor de uma variável, nós daremos um peso maior às observações obtidas em condições mais semelhantes do que naquelas obtidas em condições menos semelhantes (Gilboa *et al.*, 2011). Ou seja, com base nos dados observados, em que uma variável resposta de interesse é associada a algumas covariáveis, é possível fazer previsões usando uma média ponderada das variáveis resposta observadas, cujos pesos dependem da similaridade entre as covariáveis correspondentes.

No modelo de similaridade empírica, não há suposição sobre a existência de uma forma funcional relacionando a variável resposta e as covariáveis. O método propõe fixar uma forma particular para a função de similaridade, estimar seus parâmetros a partir dos dados e utilizar tal função para predizer a variável de interesse (Sanchez *et al.*, 2019).

Fundamentado nestes conceitos, a principal colaboração deste trabalho reside na proposta de uma adaptação ao estimador de Kaplan-Meier, em que incorporamos uma função de similaridade, cujos parâmetros são obtidos a partir do método de máxima verossimilhança adaptado à análise de sobrevivência (Zhou, 2019). Desse modo, por meio da função de similaridade, torna-se viável incluir o efeito das covariáveis ao estimar as funções de sobrevivência para cada indivíduo.

A este estimador ajustado, atribuímos a nome de Similarity-based Kaplan-Meier (SBKM). Ao longo deste estudo, dedicamos os nossos esforços à exploração de diversas propriedades do SBKM, analisando seu desempenho e suas previsões em diversas configurações e contextos distintos, por meio de cálculos numéricos conduzidos computacionalmente. Nosso principal objetivo é avaliar a capacidade preditiva do SBKM em comparação com métodos estatísticos e técnicas de aprendizado de máquina.

Dito isto, este trabalho foi estruturado para apresentar os conceitos teóricos fundamentais necessários para uma compreensão abrangente da nossa proposta. No Capítulo 2, realizamos uma revisão sobre a análise de sobrevivência, elucidando os principais termos utilizados, explicando as funções essenciais e aprofundando-nos no principal objeto de estudo desta pesquisa: o estimador de Kaplan-Meier. No Capítulo 3, discutimos a modelagem por similaridade empírica, delineando suas especificidades e as convenções adotadas nesta dissertação. No Capítulo 4, apresentamos a principal contribuição desta dissertação, que consiste na proposta de uma nova adaptação para o estimador de Kaplan-Meier, incorporando uma função de similaridade empírica à sua fórmula original. O Capítulo 5 fornece detalhes sobre os dados utilizados, a metodologia aplicada e as métricas empregadas na simulação computacional. Em seguida, no Capítulo 6 temos a exposição dos principais resultados numéricos obtidos a partir da aplicação do estimador proposto. Finalmente, as conclusões e sugestões para pesquisas futuras são abordadas no Capítulo 7.

2 ANÁLISE DE SOBREVIVÊNCIA

A análise de sobrevivência é uma área da estatística que está interessada em mensurar o tempo de duração até a ocorrência de um evento e avaliar os fatores que aumentam o risco deste acontecimento, ou seja, a variável resposta é, geralmente, o tempo até a ocorrência de um evento de interesse. Este tempo é comumente denominado na literatura como "tempo de falha" e pode ser, por exemplo, o tempo até a morte de um paciente, o tempo até o recebimento de alta, ou o tempo até inadimplência de um cliente. Apesar de sua conotação negativa, o tempo de falha não implica necessariamente em um evento indesejado ou prejudicial.

Em um experimento, apenas alguns indivíduos experienciam o evento e, posteriormente, os tempos exatos de ocorrência do evento serão desconhecidos para um subconjunto do grupo de estudo. Desse modo, a principal característica dos dados de sobrevivência é a presença de censura, definida como a "ausência da ocorrência do evento no tempo de análise" e, portanto, o tempo real para o evento é desconhecido. Nestas amostras, temos apenas a observação incompleta ou parcial da resposta. Em geral, isto significa que toda informação referente ao evento de interesse se resume ao conhecimento de que o tempo de falha é superior àquele observado.

Na análise de sobrevivência, dados censurados não tem o mesmo significado de dados faltantes/omissos. Os participantes cujos dados são censurados não são excluídos e contribuem para análise até o último intervalo em que estavam presentes. Portanto, métodos de imputação não são uma alternativa a considerar (Ferreira; Patino, 2016).

Neste contexto, técnicas estatísticas usuais, como, por exemplo, análise de regressão, não devem ser utilizadas, pois são projetadas para dados completos, em outros termos, para dados em que todas as instâncias são observadas sem qualquer restrição ou censura. Suponha, por exemplo, que um estudo tenha o interesse de comparar o tempo médio de vida de três grupos de pacientes. Se não houvesse censuras, poder-se-ia usar as técnicas de análise de variância para fazer tal comparação. No entanto, o provável é que existam censuras e deste modo, seria necessário desconsiderar todas as amostras censuradas, o que resultaria na perda de informações importantes e estimativas imprecisas ou enviesadas. Portanto, para lidar com dados censurados, faz-se necessário o uso de métodos de análise de sobrevivência que possibilitam incorporar, na análise estatística, a informação contida nos dados censurados.

Para aprofundar os conceitos teóricos desta seção, foi utilizado como referência principal o livro de "Análise de Sobrevivência Aplicada" (Colosimo; Giolo, 2006).

2.1 Características dos Dados de Sobrevivência

Os dados de sobrevivência se destacam por apresentarem informações sobre os tempos de ocorrência de eventos, frequentemente acompanhados por dados de censura. A variável resposta é função destes dois componentes. Muitas vezes, realiza-se também a medição de um conjunto de covariáveis para cada indivíduo. O tempo de falha é definido pelos seguintes elementos: o tempo inicial, a escala de medida e o evento de interesse (falha) (Colosimo; Giolo, 2006). Em um estudo, é muito importante estabelecer clara e inequivocamente esses três elementos desde o início.

2.1.1 Tempo de Falha

O tempo inicial refere-se ao tempo de início do estudo e deve ser precisamente definido. Deve-se ter cuidado para não confundir o tempo inicial com a data que o indivíduo entra no estudo, pois durante o período pré-determinado para o estudo pode-se observar diferentes indivíduos com diferentes datas de inclusão no estudo. Além disso, todos os indivíduos devem ser comparáveis na origem do estudo, com exceção de diferenças medidas pelas covariáveis. Isto significa que ao iniciar o estudo, os participantes devem ser similares em termos dos fatores que podem influenciar os resultados. O tempo inicial pode ser estabelecido, por exemplo, como a data da primeira consulta, a primeira utilização de um determinado equipamento, a data do início do tratamento de alguma doença.

A escala de medida geralmente adotada é o tempo real de observação, embora existam outras alternativas. Pode-se definir estudos onde o tempo observado é o número de anos que o indivíduo levou até a ocorrência do evento de interesse, ou ainda o número de meses, semanas, etc. Em testes de Engenharia, por exemplo, podem surgir outras escalas de medida distintas, como o número de ciclos, quilometragem ou outras medidas de carga.

O terceiro elemento é o evento de interesse. Este evento deve ser definido antes de começar o estudo e é crucial determinar de maneira precisa e objetiva o que vem a ser a falha. Estes eventos podem ser indesejáveis, como a morte de um paciente, a falha de um equipamento ou a evasão de um cliente. No entanto, nem todos são negativos; por exemplo, o paciente pode receber alta hospitalar. Além disso, em alguns estudos pode-se estar interessado em múltiplas falhas. Por exemplo, em um estudo sobre previdência social, pode-se estar interessado simultaneamente nos eventos aposentadoria programada, aposentadoria por invalidez ou morte

dos indivíduos. Apenas serão abordados neste trabalho estudos com um único tipo de falha. Sendo assim, o tempo de falha vai do tempo inicial até a ocorrência do evento medido na escala desejável.

O evento de interesse (falha) também pode ocorrer devido a uma única causa ou devido a duas, ou mais. Situações em que causas de falha competem entre si são chamadas na literatura de "riscos competitivos" (Prentice *et al.*, 1978).

2.1.2 Censura

Se para um indivíduo o evento não ocorre durante o tempo de observação do estudo, ele será descrito como censurado. Estas observações, denominadas censuras, são recorrentes em estudos de sobrevivência e podem ocorrer por diferentes razões, dentre elas, a perda de acompanhamento do indivíduo no decorrer do estudo, a ausência de falha até o término do experimento.

Neste ponto, é importante ressaltar que, mesmo censurados, todos os resultados provenientes de um estudo de sobrevivência devem ser usados na análise estatística. Em primeiro lugar, mesmo se tratando de observações incompletas, estas fornecem informações relevantes sobre o tempo de vida dos indivíduos. Além disso, a omissão das censuras no cálculo das estatísticas pode acarretar conclusões enviesadas.

Por exemplo, suponha que estamos estudando a vida útil de um componente eletrônico e coletamos dados de falha de 50 componentes, mas 10 deles continuam em operação no momento em que os dados são coletados e, portanto, sua vida útil é censurada. Se tentarmos estimar a média da vida útil desses componentes usando a abordagem tradicional de média aritmética, teríamos uma estimativa enviesada e subestimada da vida útil média, porque estaríamos ignorando a informação de censura.

Existem diferentes mecanismos de censura. Censuras do tipo I são aquelas que ocorrem em estudos que serão terminados após um período pré-estabelecido de tempo. Logo, ao término do estudo, alguns indivíduos ainda não experienciaram o evento. Censuras do tipo II resultam daqueles estudos que são finalizados após a ocorrência do evento de interesse em um número pré-estabelecido de indivíduos. Um terceiro mecanismo de censura, o do tipo aleatório, é o que mais ocorre na prática. Isto acontece quando o indivíduo é retirado do experimento no decorrer do estudo sem ter ocorrido a falha, ou também, por exemplo, se um paciente em um estudo clínico morre por uma razão diferente da estudada (Colosimo; Giolo, 2006).

Utiliza-se uma representação simples do mecanismo de censura aleatória mediante duas variáveis aleatórias. Considere T_j uma variável aleatória representando o tempo de falha de um indivíduo e seja C_j outra variável aleatória independente de T_j, representando o tempo de censura associado a este mesmo sujeito. A variável δ_j representa o indicador de censura. Os dados observados serão:

$$t_j = min(T_j, C_j) \tag{2.1}$$

e

$$\delta_j = \begin{cases} 1, & T_j \leq C_j \\ 0, & T_j > C_j. \end{cases} \tag{2.2}$$

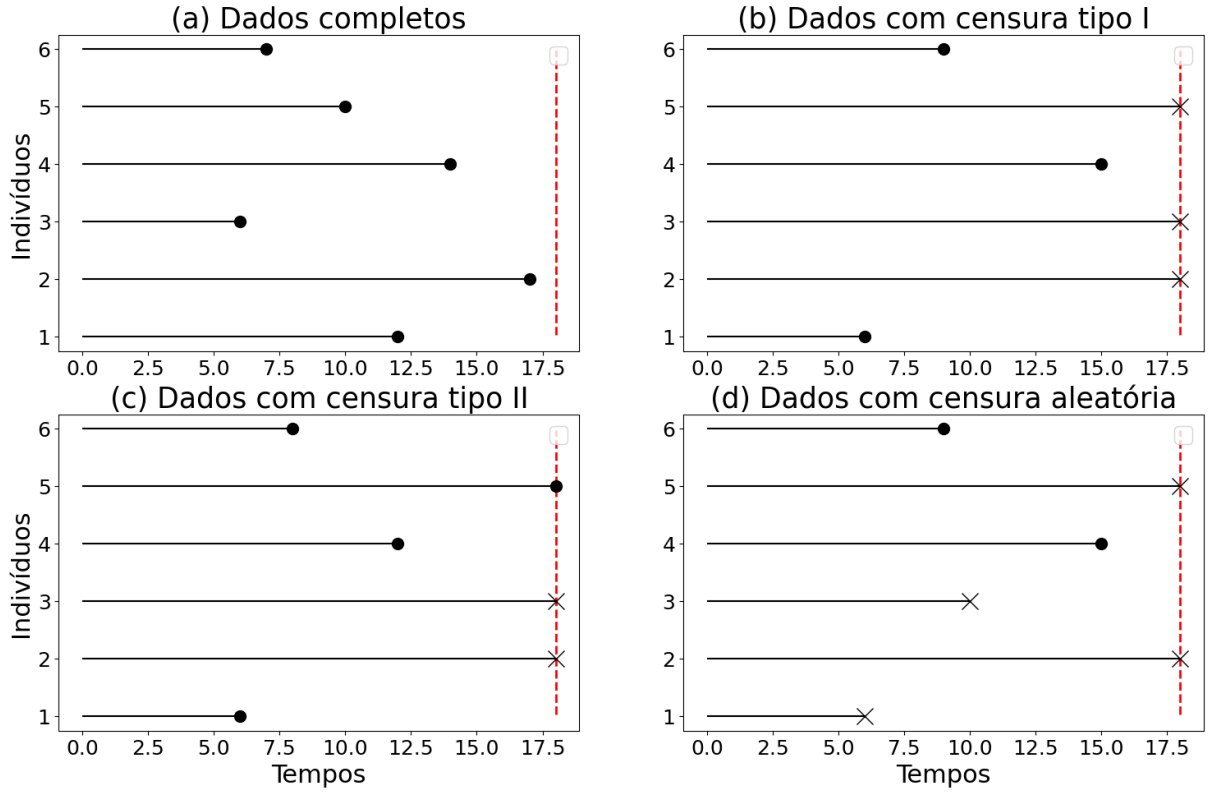
Suponha que os pares (T_j, C_j), $j = 1, ..., n$, formam uma amostra aleatória, ou seja, são vetores aleatórios bidimensionais independentes e identicamente distribuídos. Observe que se todos $C_j = C$, uma constante fixa, tem-se a censura do tipo I. Isto significa que a censura do tipo I é um caso particular da aleatória (Colosimo; Giolo, 2006). Para a censura do tipo II, temos que $C_j = C$ para todas as observações censuradas, porém C é uma variável aleatória. Por fim, na censura do tipo aleatória, os valores de C_j podem ser distintos. Dessa forma, t_j é uma variável aleatória mista pois pode assumir tanto valores contínuos quanto valores discretos.

Os mecanismos de censura ilustrados na Figura 1 são chamados censura à direita, pois o tempo de ocorrência do evento de interesse está à direita do tempo registrado. Existem outras duas formas de censura: censura à esquerda e intervalar.

A censura à esquerda ocorre quando o tempo registrado é maior que o tempo de falha. Isto acontece quando o evento de interesse já aconteceu quando o indivíduo foi observado. Por exemplo, um estudo para determinar a idade em que as crianças aprendem a ler em uma determinada comunidade, no início da pesquisa algumas crianças já sabiam ler e não se sabe com que idade isto tinha acontecido, caracterizando assim observações censuradas à esquerda (Colosimo; Giolo, 2006).

Por outro lado, a censura intervalar é um tipo mais geral de censura que acontece quando os tempos de sobrevivência não são exatamente conhecidos, sabe-se apenas que eles ocorreram em um intervalo (Strapasson, 2007). Um exemplo ilustrativo de censura intervalar são estudos nos quais os pacientes são monitorados em visitas periódicas, e a única informação disponível é que o evento ocorreu em um determinado intervalo de tempo.

Figura 1 – Mecanismos de censura



Fonte: Elaborado pela autora.
Nota: Ilustração de alguns mecanismos de censura em que • representa falha e × representa censura: (a) todos os indivíduos experimentaram o evento antes do final do estudo, (b) os indivíduos 2, 3 e 5 não experimentaram o evento até o final do estudo, (c) o estudo foi finalizado após a ocorrência de um número pré-estabelecido de 4 falhas e (d) o acompanhamento de alguns indivíduos foi interrompido por alguma razão e outros indivíduos não experimentaram o evento até o final do estudo.

Entretanto, neste estudo, restringiremos nossa análise a dados de sobrevivência com censura à direita, uma situação comumente encontrada em campos como Medicina, Engenharia e Ciências Sociais. Logo, ao mencionarmos a palavra "censura" neste trabalho, estaremos nos referindo especificamente às censuras à direita.

2.2 Representação dos Dados de Sobrevivência

Os dados de sobrevivência para o indivíduo j ($j = 1, \ldots, n$) sob estudo são representados pelo par (t_j, δ_j) sendo t_j o tempo de falha ou de censura e δ_j a variável indicadora de falha ou censura, isto é,

$$\delta_j = \begin{cases} 1, & \text{se } t_j \text{ é um tempo de falha} \\ 0, & \text{se } t_j \text{ é um tempo censurado.} \end{cases} \tag{2.3}$$

Assim, a variável aleatória resposta em análise de sobrevivência é representada por duas colunas no conjunto de dados.

Quando existem covariáveis medidas no j-ésimo indivíduo, como, por exemplo, $\mathbf{x}_j$ = (sexo$_j$, idade$_j$, estágio da doença$_j$, etc), os dados serão representados por $(t_j, \delta_j, \mathbf{x}_j)$. Em especial, quando temos dados de sobrevivência intervalar, tem-se, ainda, a representação $(l_j, u_j, \delta_j, \mathbf{x}_j)$ em que l_j é o limite inferior e u_j o limite superior do intervalo observado para o j-ésimo participante.

2.3 Tempo de Sobrevivência

Os dados de sobrevivência são geralmente descritos e modelados em termos de duas funções relacionadas, sobrevivência e risco. Desse modo, a variável aleatória não-negativa T, que representa o tempo de falha, é especificada pela sua função de sobrevivência ou pela sua função de risco.

2.3.1 Função de Sobrevivência

A função de sobrevivência $S(t)$ é definida como a probabilidade de uma observação não falhar até um certo tempo t, ou seja, a probabilidade de uma observação sobreviver ao tempo t. Isto é descrito como

$$S(t) = P(T > t). \tag{2.4}$$

Como consequência, a função de distribuição acumulada $F(t)$ é definida como a probabilidade de uma observação não sobreviver ao tempo t

$$F(t) = 1 - S(t) = P(T \leq t). \tag{2.5}$$

Algumas características da função de sobrevivência:

1. Toda função de sobrevivência $S(t)$ é monotonicamente não crescente, ou seja, $S(u) \leq S(t)$ $\forall u > t$;
2. S(t) é contínua à direita, isto é, $\lim_{t \to t_0^+} S(t) = S(t_0)$;
3. Em $t = 0$, $S(t) = 1$;
4. $\lim_{t \to \infty} S(t) = 0$, isto significa que $S(t)$ se aproxima assintoticamente de zero à medida que t tende ao infinito.

2.3.2 Função de Risco

A função de risco, ou taxa de falha, é geralmente denotada por $h(t)$ e representa a taxa de falha instantânea no tempo t condicional à sobrevivência até o tempo t. Formalmente, a função de risco de T, em que T é uma variável aleatória absolutamente contínua não negativa, é definida como

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t < T < t + \Delta t | T > t)}{\Delta t} = \frac{f(t)}{S(t)} = -\frac{d}{dt} \log S(t). \tag{2.6}$$

Observe que, em contraste com a função de sobrevivência, que se concentra em não ter um evento, a função de risco se concentra na ocorrência do evento. Em resumo, o risco está relacionado à taxa de eventos incidentes, enquanto a sobrevivência reflete a não ocorrência cumulativa.

As taxas de falha são sempre positivas e não têm limite superior definido. Uma função de risco crescente indica que a probabilidade de falha de um indivíduo aumenta com o tempo, enquanto uma função decrescente indica o oposto.

2.3.3 Tempo Médio

Uma quantidade de interesse em análise de sobrevivência é o tempo médio de vida. O tempo médio pode ser obtido pela área sob a curva de sobrevivência. Isto é,

$$t_m = \int_0^\infty S(t) dt \tag{2.7}$$

O tempo médio de vida refere-se ao tempo de sobrevivência esperado que um indivíduo possui antes da ocorrência do evento de interesse, ou seja, antes de falhar. Ou seja, a Equação 2.7 representa o valor esperado de T, obtido diretamente da função de sobrevivência, sem necessidade de estimar densidade ou função de probabilidade.

2.4 Estimador de Kaplan-Meier

Uma das principais componentes na análise de dados envolvendo tempos de vida é a função de sobrevivência. Dito isso, precisamos encontrar uma estimativa para a função de sobrevivência e, então, a partir dela, calcular as estatísticas de interesse, como, por exemplo, tempo médio ou mediano.

O estimador de Kaplan-Meier, também chamado de estimador limite-produto, é um método não paramétrico, proposto por Kaplan e Meier (1958), usado para estimar a função de sobrevivência a partir dos tempos de sobrevivência observados, censurados e não censurados. Este estimador é uma adaptação da função de sobrevivência empírica que, na ausência de censuras, é definida como:

$$\widehat{S}(t) = \frac{1}{\mathrm{n}} \sum_{j=1}^{n} \mathbb{1}\{t_j > t\}. \tag{2.8}$$

A função de sobrevivência empírica representa um estimador não viesado, o que significa que, para uma amostra suficientemente grande, ela tende a ser uma estimativa correta da verdadeira função de sobrevivência da população. Isto é fundamentado no teorema de Glivenko-Cantelli, que afirma que a função de distribuição empírica converge uniformemente para a verdadeira função de distribuição à medida que o tamanho da amostra tende ao infinito (Salnikov, 2021). Além disso, a função de sobrevivência empírica não apenas converge para a função verdadeira à medida que a amostra cresce, mas essa convergência ocorre de forma uniforme em todos os pontos do tempo t, ou seja, é um estimador uniformemente consistente (Gill, 1994).

$\widehat{S}(t)$ assume a forma de uma função escada com degraus localizados nos tempos observados de falha, sendo o tamanho de cada degrau igual a $1/n$, em que n é o tamanho da amostra. Em caso de empates em um certo tempo t, o tamanho do degrau é multiplicado pelo número de empates.

O estimador de Kaplan-Meier, ao ser construído, contempla um número de intervalos de tempo equivalente ao total de falhas distintas. Os limites desses intervalos são determinados pelos tempos de falha observados na amostra.

Considere que n indivíduos sob estudo no período de acompanhamento têm falhas ou censuras em tempos distintos t_i, onde $i \in \{1,2,\ldots,n'\}$ e $n' \leq n$. Aqui, t_i representa o i-ésimo menor valor no conjunto $\{t_1,t_2,\ldots,t_n\}$. Como se supõe que os eventos ocorram de forma independente entre si, as probabilidades de um indivíduo qualquer sobreviver de um intervalo para o seguinte podem ser multiplicadas para obter a probabilidade de sobrevivência cumulativa (Clark *et al.*, 2003). A probabilidade de um indivíduo estar vivo em um tempo t é calculada por

$$\widehat{S}(t) = \prod_{i=1}^{n'} \left(1 - \frac{d_i}{g_i}\right)^{\delta_i \mathbb{1}\{t_i \leq t\}}, \tag{2.9}$$

em que $d_i = \sum_{j=1}^{n} \delta_j \mathbb{1}\{t_j = t_i\}$ é o número de falhas em t_i e $g_i = \sum_{j=1}^{n} \mathbb{1}\{t_j \geq t_i\}$ é o número de

indivíduos sob risco em t_i, isto é, os indivíduos que não falharam e não foram censurados até o instante imediatamente anterior a t_i.

Naturalmente, o estimador de Kaplan-Meier se reduz à função de sobrevivência empírica (2.8) se não houver censuras. Além disso, este estimador não atinge o valor $\widehat{S}(t) = 0$ quando o maior tempo observado na amostra corresponder a uma censura. Isto acontece em estudos envolvendo os mecanismos de censura do tipo I e II, uma vez que as últimas observações costumam ser censuradas (Colosimo; Giolo, 2006).

No artigo original, Kaplan e Meier justificam a expressão (2.9) mostrando que ela é o estimador de máxima verossimilhança de $S(t)$ (Kaplan; Meier, 1958). Suponha, como feito anteriormente, que temos n observações que falham ou são censuradas no tempo t_j, para $j = 1, ..., n$. Para observações não censuradas, a probabilidade de falha no tempo t_j pode ser calculada da seguinte forma:

$$S(t_j^-) - S(t_j), \tag{2.10}$$

com $S(t_j^-) = \lim_{\Delta t \to 0} S(t_j - \Delta t)$. Por outro lado, a contribuição para a função de verossimilhança de um tempo de sobrevivência censurado em t_i é:

$$P(T > t_j) = S(t_j). \tag{2.11}$$

A função de verossimilhança pode, então, ser escrita como:

$$L(S(\cdot)) = \prod_{j=1}^{n} \left[S(t_j^-) - S(t_j)\right]^{\delta_j} \left[S(t_j)\right]^{1-\delta_j}. \tag{2.12}$$

Pode-se mostrar que $S(t)$ que maximiza $L(S(\cdot))$ é exatamente a expressão (2.9). Esta definição do estimador de máxima verossimilhança é uma generalização do conceito convencional empregado em modelos paramétricos. Como o estimador é não-paramétrico, trata-se, na verdade, de uma estimação por máxima quase-verossimilhança (Gourieroux *et al.*, 1984), pois não temos a verossimilhança exata.

Ademais, para avaliar a precisão e testar hipóteses para $\widehat{S}(t)$, faz-se necessário definir os intervalos de confiança para o estimador. A expressão para a variância assintótica do estimador de Kaplan-Meier é dada por:

$$\text{Var}(\widehat{S}(t)) = \left[\widehat{S}(t)\right]^2 \sum_{i=1}^{n'} \left[\frac{d_i}{g_i(g_i - d_i)}\right] \delta_i \mathbb{1}\{t_i \leq t\}. \tag{2.13}$$

Esta expressão é conhecida como fórmula de Greenwood e pode ser obtida a partir de propriedades do estimador de máxima verossimilhança (Kalbfleisch; Prentice, 2002). Como $\widehat{S}(t)$, para

t fixo, tem distribuição assintótica Normal, segue que, em um intervalo de $100(1-\alpha)\%$ de confiança assintótico para $S(t)$, é dado por:

$$\left[\widehat{S}(t) \pm z_{\alpha/2}\sqrt{\text{Var}(\widehat{S}(t))}\right], \tag{2.14}$$

em que $z_{\alpha/2}$ denota o $\alpha/2$-percentil superior da distribuição Normal padrão.

Para valores extremos de t, este intervalo de confiança pode apresentar limite inferior negativo ou limite superior maior do que 1. Nesses casos, o problema é resolvido utilizando uma transformação para $S(t)$ como, por exemplo, $\widehat{U}(t) = \log[-\log(\widehat{S}(t))]$ que tem variância assintótica estimada por

$$\text{Var}(\widehat{U}(t)) = \frac{\sum_{i=1}^{n'}\left[\frac{d_i}{g_i(g_i-d_i)}\right]\delta_i \mathbb{1}\{t_i \leq t\}}{\sum_{i=1}^{n'}\left[\log\left(\frac{g_i-d_i}{g_i}\right)\right]\delta_i \mathbb{1}\{t_i \leq t\}}. \tag{2.15}$$

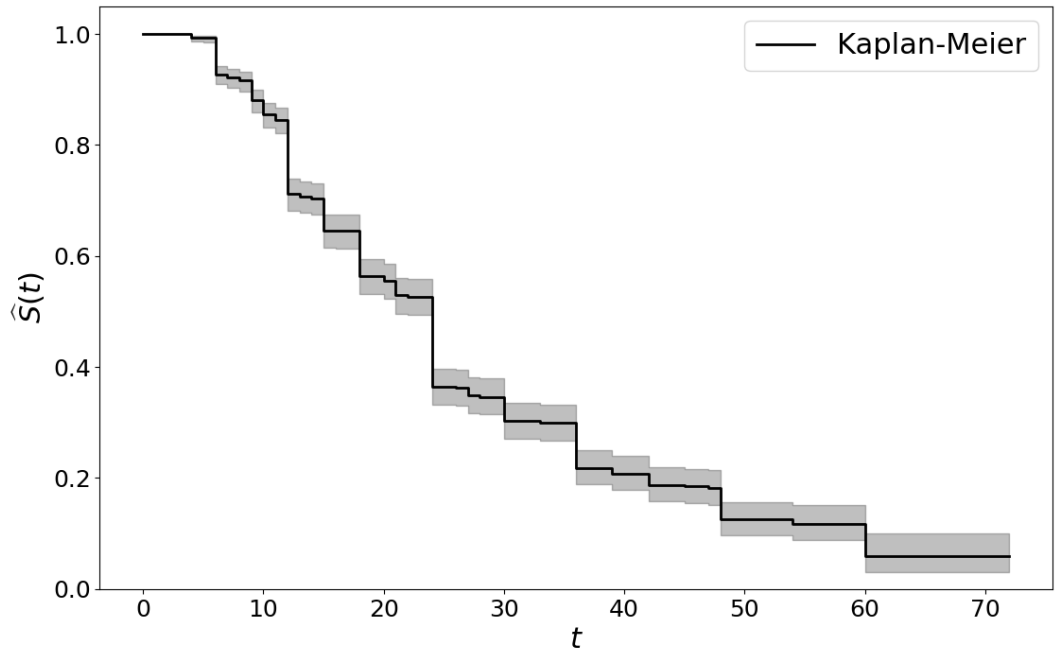
Assim, um intervalo aproximado de $100(1-\alpha)\%$ de confiança para $S(t)$ é dado por:

$$\left[\widehat{S}(t)^{\exp\{\pm z_{\alpha/2}\sqrt{\text{Var}(\widehat{U}(t))}\}}\right], \tag{2.16}$$

que assume valores no intervalo $[0,1]$ (Colosimo; Giolo, 2006).

Nota-se que a representação gráfica do estimador de Kaplan-Meier é construída mantendo o valor de $\widehat{S}(t)$ constante entre os tempos de falha. Para ilustração, a curva de sobrevivência estimada, juntamente com os intervalos de confiança, é apresentada na Figura 2 para o conjunto de dados que será descrito na Seção 5.1. Vale ressaltar que, conforme mencionado, este gráfico não atinge $\widehat{S}(t) = 0$, pois o maior tempo observado na amostra é uma censura.

Figura 2 – Estimativa de Kaplan-Meier para os dados de risco de crédito apresentados na Seção 5.1 elaborada utilizando a biblioteca *lifelines* disponível em Python (Davidson-Pilon, 2019)



Fonte: Elaborada pela autora.

A curva de sobrevivência na Figura 2, fornece um resumo útil dos dados que pode ser usado para estimar medidas como o tempo médio de sobrevivência. Porém, por conta da distorção encontrada na distribuição da maioria dos dados sobrevivência, a média não é frequentemente usada (Clark *et al.*, 2003).

O estimador de Kaplan-Meier não é capaz de realizar estimativas com preditores quantitativos, como, por exemplo, a idade. Deste modo, as curvas de Kaplan–Meier são mais úteis quando as covariáveis de interesse são categóricas (por exemplo, sexo) ou assumem um número pequeno de valores que podem ser tratados como categorias.

O teste de *logrank* desempenha um papel crucial como um teste estatístico não-paramétrico em estudos de sobrevivência, visando determinar se existem diferenças significativas na sobrevivência entre dois ou mais grupos (Mantel, 1966). Sua aplicação envolve a comparação das estimativas de KM das funções de sobrevivência dos grupos em questão, utilizando um teste de hipótese para verificar a significância estatística da diferença entre as curvas. Este teste é particularmente apropriado quando a razão das funções de taxa de falha dos grupos a serem comparados é aproximadamente constante. Ou seja, as populações têm a propriedade de riscos proporcionais.

3 SIMILARIDADE EMPÍRICA

Frequentemente, estamos interessados em determinar o valor de uma determinada variável específica de interesse e, em diversas situações, os dados disponíveis são pertinentes ao problema, mas não sugerem uma estimativa razoável a ser feita. A similaridade, enquanto aspecto subjacente a muitos modelos e métodos estatísticos, proporciona uma abordagem para resolver esse dilema. Se dois objetos, como pessoas, empresas ou produtos, são similares em alguns aspectos, podemos inferir que compartilham outras características em comum.

Por exemplo, considere que uma pessoa deseja vender seu apartamento na cidade e se questiona como determinar o valor do imóvel. Provavelmente a decisão do preço de venda pode ser baseada em informações disponíveis do apartamento como uma função das suas características tais como localização, área, vista, e assim por diante. Desta forma, o valor do apartamento pode ser visto como um modelo simplificado que o dono pode usar como base para determinar o valor da venda. Dada uma base de dados com o valor de venda de outros apartamentos, como o dono poderia avaliar o valor de seu apartamento? É razoável sugerir que as decisões devem ser tomadas por analogias com casos passados em situações semelhantes.

Neste problema, tentamos avaliar o valor de uma variável y_j baseado nos valores das covariáveis $\mathbf{x}_j = (x_j^1, ..., x_j^m)^\intercal$, e na informação contida em uma base de dados composta de variáveis $(x_i^1, ..., x_i^m, y_i)^\intercal$, para todo $i = 1, ..., n$.

O método de similaridade empírica propõe combinar observações passadas de $\mathbf{x}$ e y com os valores atuais de $\mathbf{x}$ para gerar uma avaliação de y utilizando a média ponderada por similaridade. Isto significa que o valor predito de y, $\widehat{y}_j$, será a média ponderada de todos os valores y_i observados anteriormente, onde o peso de y_i, para todo $i = 1, ..., n$, é a similaridade entre o vetor $(x_j^1, ..., x_j^m)^\intercal$, associado a y_j, e o vetor observado anteriormente $(x_i^1, ..., x_i^m)^\intercal$. Esta função de similaridade não tem uma forma funcional particular e deve ser fixada conforme o contexto da pesquisa e seus parâmetros estimados a partir dos dados para prever a variável resposta (Gilboa *et al.*, 2006).

O preditor baseado em similaridade de y_j, dada uma função de similaridade s, é definido como:

$$\widehat{y}_j = \frac{\sum_{i \neq j} s(\mathbf{x}_i, \mathbf{x}_j) y_i}{\sum_{i \neq j} s(\mathbf{x}_i, \mathbf{x}_j)}. \tag{3.1}$$

A similaridade empírica pode ser vista como uma estimativa pontual de uma função de similaridade se incorporarmos a equação (3.1) em um modelo estatístico. Esta abordagem

para estimar valores previstos evoca a lembrança do método de estimativa por kernel (Akaike, 1954).

A princípio, estamos interessados em funções de similaridade que dependam de uma distância ponderada. Considere $\mathbf{w} \in \mathbb{R}_+^m$ o vetor de pesos tal que a distância de dois vetores $\mathbf{x}_i$, $\mathbf{x}_j$ seja dada pelo quadrado da Distância Euclidiana Ponderada (DEP):

$$d_w(\mathbf{x}_i, \mathbf{x}_j) = \sum_{l=1}^{m} w_l (x_i^l - x_j^l)^2. \tag{3.2}$$

A principal motivação para optar pela distância euclidiana ponderada em vez da distância euclidiana convencional reside na consideração de que diferentes covariáveis podem exercer influências distintas sobre a variável resposta. Por exemplo, ao analisar o valor de compra e venda de um imóvel, o número de quartos pode ser mais significativo do que a área total.

Em seguida, o objetivo é migrar de uma função de distância para uma função de similaridade. Esta última deve ser decrescente na distância d_w, atingir o valor de 1 quando $d_w = 0$, e convergir para 0 à medida que $d_w \to \infty$. Possíveis candidatos naturais para essa função incluem

$$s_w(\mathbf{x}_i, \mathbf{x}_j) = \begin{cases} e^{-d_w(\mathbf{x}_i, \mathbf{x}_j)} & \text{(EX)}, \\ \\ \frac{1}{1+d_w(\mathbf{x}_i, \mathbf{x}_j)} & \text{(FR)}, \end{cases} \tag{3.3}$$

em que serão usadas funções de similaridade do tipo exponencial (EX) e fracionário (FR). Observe que, inicialmente, os pesos w_l não têm a restrição de que a sua soma seja igual a 1, o que implica uma maior flexibilidade nos pesos das observações mais próximas em comparação com aquelas mais distantes na base de dados (Sanchez, 2015).

No entanto, uma "função de similaridade empírica" pode ser qualquer função escolhida para ajustar os dados segundo a Equação (3.1). A essência da similaridade empírica é justamente que uma função de similaridade seja fixada e seus parâmetros estimados a partir dos mesmos dados que são utilizados, em conjunto com esta função, para realizar previsões (Gilboa *et al.*, 2011).

3.1 Covariáveis quantitativas

No modelo de similaridade é possível gerar predições utilizando a expressão (3.1). Assim, para cada valor predito $\widehat{y}_j$ o modelo usa todas as observações, em uma expressão que

pode ser vista como um interpolador linear local. Esta predição é semelhante ao estimador de Nadaraya-Watson para regressão não paramétrica, em que $s(\mathbf{x}_i, \mathbf{x}_j)/\sum_{i\neq j} s(\mathbf{x}_i, \mathbf{x}_j)$ desempenha o papel da função de kernel (Sanchez, 2015).

Observamos que em determinadas aplicações, as covariáveis observadas são de natureza categórica. Nesse contexto, o método de estimação referente à equação (3.1) não pode ser adotado diretamente, uma vez que o cálculo da equação (3.2) é inviável, dado que a DEP está restrita a covariáveis numéricas.

Dessa forma, atribuiremos à distância (3.2) a designação d_w^Q, incorporando o sobrescrito Q para indicar explicitamente que esse cálculo é destinado apenas à covariáveis quantitativas.

3.2 Covariáveis categóricas nominais

A metodologia para o cálculo da função de similaridade entre covariáveis quantitativas é simples e intuitivo. Porém, em muitas situações práticas envolvendo conjuntos de dados de alta dimensão ou estruturados, nem todas covariáveis são numéricas.

Quando existem covariáveis qualitativas disponíveis, a noção de distância euclidiana dada na expressão (3.2) não está bem estabelecida e algumas adaptações são necessárias. Em (Gayer *et al.*, 2007), os autores sugerem codificar uma variável categórica em variáveis indicadoras do tipo *dummy*, que são variáveis binárias (0 ou 1) criadas para representar uma variável com duas ou mais categorias, e só então aplicar a distância euclidiana ponderada para avaliar a similaridade.

Assim, cada covariável indicadora teria um peso distinto a ser estimado, independentemente de pertencerem todas à mesma covariável qualitativa original ou não. Consequentemente, embora representem diferentes níveis da mesma covariável, uma covariável indicadora pode ter uma influência maior na estimativa da variável resposta do que outra covariável indicadora. Outra desvantagem deste método reside no aumento da dimensionalidade dos dados, o qual pode ser significativo dependendo do número de categorias em cada covariável.

Entretanto, podemos utilizar uma abordagem alternativa que não codifica as variáveis qualitativas em variáveis *dummy*. Em vez disso, Sanchez *et al.* (2019) propõe medir distâncias de valores observados de covariáveis categóricas através de uma distância binária ponderada que apenas distingue se os valores observados são iguais ou não. Com essa abordagem, diferentes níveis de uma covariável qualitativa são sempre associados a um mesmo peso, o que, além de

reduzir o número de parâmetros a serem estimados, parece ser mais congruente e de mais fácil interpretação.

Para isso, faremos o uso da seguinte métrica de Distância Binária Ponderada (DBP):

$$d_w^{CN}(\mathbf{x}_i, \mathbf{x}_j) = \sum_{l=1}^{m} w_l I(x_i^l, x_j^l), \tag{3.4}$$

em que $\mathbf{x}_i = (x_i^1, ..., x_i^m)^\intercal$ e $\mathbf{x}_j = (x_j^1, ..., x_j^m)^\intercal$ são duas observações de m covariáveis categóricas nominais (*CN*), e

$$I(x_i^l, x_j^l) = \begin{cases} 0, & \text{se } x_i^l = x_j^l \\ 1, & \text{se } x_i^l \neq x_j^l. \end{cases} \tag{3.5}$$

A distância data em (3.4) nada mais é que uma distância de Hamming Ponderada (Wegner, 1960). Desta maneira, a distância será dada pela soma dos pesos cujas covariáveis são diferentes. Ou seja, as predições da variável de interesse serão a média ponderada dos valores passados da variável de interesse, onde as que possuírem mais características em comum terão um maior peso. Esta alternativa equivale a codificar as covariáveis em variáveis *dummy*, mas limitando os pesos a serem iguais para todos os níveis da mesma covariável sem aumentar a dimensionalidade da matriz modelo.

3.3 Covariáveis categóricas ordinais

Embora a distância binária sirva para codificar a similaridade entre variáveis categóricas nominais, ela não é adequada para variáveis categóricas ordinais. Isto ocorre porque essa distância não consegue refletir a relação de ordem original das categorias. Entretanto, um conjunto numérico pode ser utilizado para substituir o conjunto original de valores associados a cada categoria, desde que a ordem original das categorias seja mantida. Chamamos essa transformação de transformação monotônica. Uma transformação monotônica é uma função matemática que preserva a ordem dos valores em um conjunto de dados. Em outras palavras, quando uma transformação é monotônica, se um valor A é maior do que um valor B no conjunto de dados original, então a transformação aplicada a esses valores também indicará que o valor transformado de A é maior do que o valor transformado de B.

A hierarquia do exército é um exemplo de variável com nível ordinal de mensuração. Tomemos como exemplo as categorias Soldado, Cabo e Sargento. É importante notar que, em termos de patente, Soldado é inferior a Cabo, que por sua vez é inferior a Sargento. Assim,

é possível atribuir o valor 1 à categoria Soldado, 2 à categoria Cabo e 3 à categoria Sargento, mantendo a ordem original das patentes na nova escala numérica. No entanto, é preciso observar que apesar de que Cabo seja uma patente menor do que a de Sargento e maior do que Soldado, não necessariamente podemos supor que a diferença entre Cabo e Soldado seja igual à diferença entre Cabo e Sargento.

Portanto, os valores numéricos atribuídos às variáveis categóricas ordinais não são verdadeiramente quantitativos e sim uma aproximação que pode não capturar totalmente a estrutura dos dados, logo quaisquer análises ou conclusões tiradas deles devem ser interpretadas com cautela. Também é importante considerar o contexto e o significado das categorias ao atribuir valores numéricos a elas.

Em resumo, para calcular a distância entre as categorias de uma covariável categórica ordinal é necessário ranquear numericamente cada categoria com base em sua ordem e, em seguida, usar medidas estatísticas apropriadas. Neste trabalho, usaremos a Distância Euclidiana Normalizada (DEN) pela maior distância possível entre os ranques:

$$d_w^{CO}(\mathbf{x}_i, \mathbf{x}_j) = \sum_{l=1}^{m} \frac{w_l (x_i^l - x_j^l)^2}{(x_{max}^l - x_{min}^l)^2}, \tag{3.6}$$

em que $\mathbf{x}_i = (x_i^1, ..., x_i^m)^\intercal$ e $\mathbf{x}_j = (x_j^1, ..., x_j^m)^\intercal$ são duas observações de m covariáveis categóricas ordinais (CO).

A medida de distância resultante terá valores no intervalo $[0, 1]$, com 0 indicando vetores idênticos e 1 indicando vetores maximamente diferentes. Essa técnica de normalização pode ser útil em situações em que os valores absolutos das distâncias não são tão importantes quanto seus valores relativos ou ao comparar distâncias em conjuntos de dados com covariáveis em intervalos variados.

3.4 Caso geral

Para lidar com variados tipos de covariáveis, estabelecemos diferentes métricas de distância. Essa abordagem procura assegurar que avaliamos a similaridade entre as covariáveis de maneira apropriada, levando em consideração suas características individuais. Desse modo, torna-se necessário definir uma distância global para casos gerais, nos quais um conjunto de dados inclui covariáveis quantitativas, categóricas nominais e categóricas ordinais.

Nesse sentido, no caso geral que envolve os três tipos de covariáveis, a distância total é obtida somando as distâncias previamente definidas em (3.2), (3.4) e (3.6). Esse cálculo leva em consideração cada subconjunto de covariáveis, aplicando a devida distância correspondente.

Suponha que possuímos um conjunto de dados com m covariáveis. Definimos $m_1 + m_2 + m_3 = m$, em que m_1, m_2 e m_3 representam as quantidades de covariáveis quantitativas, categóricas nominais e categóricas ordinais, respectivamente. Dessa forma, calculamos a distância total entre as duas amostras da seguinte maneira:

$$\begin{aligned} d_w &= d_w^Q + d_w^{CN} + d_w^{CO} \\ &= \sum_{l=1}^{m_1} w_l (x_i^l - x_j^l)^2 + \sum_{l=m_1+1}^{m_1+m_2} w_l I(x_i^l, x_j^l) + \sum_{l=m_1+m_2+1}^{m} \frac{w_l (x_i^l - x_j^l)^2}{(x_{max}^l - x_{min}^l)^2}. \end{aligned} \tag{3.7}$$

4 ESTIMADOR DE KAPLAN-MEIER BASEADO EM SIMILARIDADE

Considere uma base de dados consistindo em observações da forma $(t_j, \delta_j, \mathbf{x}_j)$, para $j = 1,2,...,n$, onde $\mathbf{x}_j = (x_j^1,...,x_j^m)^\intercal$ é um vetor de m covariáveis associadas à j-ésima observação, δ_j é uma variável binária que é igual a 1 se uma falha ocorre na j-ésima observação e igual a 0, se uma censura ocorre na j-ésima observação, t_j denota os tempos ordenados de falha ou censura da j-ésima observação dependendo se $\delta_j = 1$ ou $\delta_j = 0$, respectivamente, e defina $t_0 = 0$. Os tempos distintos t_i são definidos como o i-ésimo menor valor no conjunto $\{t_1, t_2, \ldots, t_n\}$, onde $i \in 1,2,\ldots,n'$ e $n' \leq n$. Seja T variável aleatória que descreve o tempo até a falha. Para um dado argumento, considere $\mathbb{1}\{\cdot\}$ função indicadora, isto é, recebe o valor 1 quando seu argumento é verdadeiro e 0 caso contrário. Finalmente, seja $s_w(\mathbf{x}, \mathbf{x}')$ uma função de similaridade empírica pré-especificada que mede quão semelhantes são as covariáveis $\mathbf{x}$ e $\mathbf{x}'$ (onde um valor mais alto indica maior semelhança). Esta função tem como parâmetro o vetor de pesos $\mathbf{w}$, cujas componentes estão associadas a cada covariável.

Para um dado vetor de características $\mathbf{x} = (x^1,...,x^m)^\intercal$, devemos estimar a função de sobrevivência condicional $S(t|\mathbf{x}) = P(T > t|\mathbf{x})$. Neste ponto, propomos a utilização do estimador de Kaplan-Meier baseado em Similaridade (SBKM) para estimar a função de sobrevivência condicional. Essa estimativa é definida como segue:

$$\widehat{S}(t|\mathbf{x}) = \prod_{i=1}^{n'} \left[1 - \frac{\sum_{j=1}^{n} s_w(\mathbf{x}, \mathbf{x}_j) \delta_j \mathbb{1}\{t_j = t_i\}}{\sum_{j=1}^{n} s_w(\mathbf{x}, \mathbf{x}_j) \mathbb{1}\{t_j \geq t_i\}} \right]^{\delta_i \mathbb{1}\{t_i \leq t\}}. \tag{4.1}$$

Nota-se que para o caso especial em que $s_w(\mathbf{x}, \mathbf{x}_j) = 1$ para todo j, voltamos ao estimador clássico de Kaplan-Meier (KM), o qual estima a curva de sobrevivência sem o uso de covariáveis, conforme especificado em 2.9.

Para estimar os valores de cada componente de $\mathbf{w}$, podemos empregar o método de máxima verossimilhança, que busca obter a estimativa mais plausível numa amostra para o parâmetro populacional desconhecido de interesse, neste caso, o vetor de pesos $\mathbf{w}$. Especificamente, utilizaremos o método de máxima verossimilhança empírica discutido por Zhou (2019), que consiste em uma generalização do método tradicional aplicado em modelos paramétricos, conforme definido na Equação 2.12, baseado em Colosimo; Giolo (2006). A função de verossimilhança considerada é dada por:

$$L(w) = \prod_{k=1}^{n} [\widehat{P}(t_k|\mathbf{x}_k)]^{\delta_k} [\widehat{S}(t_k|\mathbf{x}_k)]^{1-\delta_k}, \tag{4.2}$$

em que

$$\widehat{S}(t_k|\mathbf{x}_k) = \prod_{i=1}^{n'} \left[1 - \frac{\sum_{j=1}^{n} s_w(\mathbf{x}_k, \mathbf{x}_j)\delta_j \mathbb{1}\{t_j = t_i\}}{\sum_{j=1}^{n} s_w(\mathbf{x}_k, \mathbf{x}_j)\mathbb{1}\{t_j \geq t_i\}} \right]^{\delta_i \mathbb{1}\{t_i \leq t_k\}} \tag{4.3}$$

é a própria curva de sobrevivência estimada pelo SBKM em t_k dado o vetor de covariáveis $\mathbf{x}_k$ e

$$\begin{aligned} & \widehat{P}(t_k|\mathbf{x}_k) = \widehat{S}(t_k^-|\mathbf{x}_k) - \widehat{S}(t_k|\mathbf{x}_k) \qquad (4.4) \\ = & \prod_{i=1}^{n'} \left[1 - \frac{\sum_{j=1}^{n} s_w(\mathbf{x}_k, \mathbf{x}_j)\delta_j \mathbb{1}\{t_j = t_i\}}{\sum_{j=1}^{n} s_w(\mathbf{x}_k, \mathbf{x}_j)\mathbb{1}\{t_j \geq t_i\}} \right]^{\delta_i \mathbb{1}\{t_i < t_k\}} - \prod_{i=1}^{n'} \left[1 - \frac{\sum_{j=1}^{n} s_w(\mathbf{x}_k, \mathbf{x}_j)\delta_j \mathbb{1}\{t_j = t_i\}}{\sum_{j=1}^{n} s_w(\mathbf{x}_k, \mathbf{x}_j)\mathbb{1}\{t_j \geq t_i\}} \right]^{\delta_i \mathbb{1}\{t_i \leq t_k\}} \\ = & \widehat{S}(t_k^-|\mathbf{x}_k) \frac{\sum_{j=1}^{n} s_w(\mathbf{x}_k, \mathbf{x}_j)\delta_j \mathbb{1}\{t_j = t_k\}}{\sum_{j=1}^{n} s_w(\mathbf{x}_k, \mathbf{x}_j)\mathbb{1}\{t_j \geq t_k\}} \end{aligned}$$

é a probabilidade de falha no tempo t_k dado o vetor de covariáveis $\mathbf{x}_k$.

O estimador SBKM, conforme definido na Equação 4.1, presume uma função de similaridade predefinida s que permanece constante tanto em relação às realizações de y_i quanto ao próprio índice de observação i. Uma vez que estabelecemos a função de similaridade, derivada das distâncias ponderadas calculadas conforme a Equação 3.7, e ao especificarmos as possíveis formas dessa função na Equação 3.3, ganhamos flexibilidade para lidar com covariáveis que possuem diferentes influências na métrica de distância. Assim, quanto mais próximo $\mathbf{x}_k$ estiver de $\mathbf{x}_j$, maior será o peso que y_k receberá de y_j, em relação aos outros $y'_j s$, na construção da previsão.

4.1 Exemplo fictício

Para ilustrar o método proposto, faremos uso de um exemplo prático utilizando dados fictícios para demonstrar o algoritmo que deverá ser implementado e testado computacionalmente posteriormente. Suponha que temos uma base de dados $(t_i, \delta_i, \mathbf{x}_i)$, para $i = 1, 2, ..., 4$, onde $\mathbf{x}_i = (x_i^1, x_i^2)^\intercal$ é um vetor de duas covariáveis associadas à i-ésima observação.

Tabela 1 – Dados fictícios

id	Tempo	Falha	Estágio	Idade
1	2.5	0	1	76
2	3.8	1	4	40
3	7.0	1	2	54
4	10.0	0	3	68
5	-	-	1	72

Fonte: Elaborada pela autora.

Esses dados representam uma simulação de um estudo envolvendo pacientes diagnosticados com uma determinada doença específica, acompanhando o tempo até o óbito do paciente em meses. Os dados consistem em um indicador de censura (Falha), a idade do paciente (Idade), representada como uma variável numérica, e o estágio da doença (Estágio), que é uma variável categórica ordinal. Observa-se que as covariáveis estão em escalas diferentes, o que pode distorcer a medida de distância devido à disparidade nas magnitudes das variáveis. Por exemplo, se uma variável está na escala de milhares e outra na escala de unidades, a distância será predominantemente influenciada pela variável de maior magnitude, comprometendo a interpretação de similaridade entre os pontos. Para assegurar que todas as covariáveis contribuam igualmente para a medida de distância, optamos por padronizar a covariável numérica Idade. A variável Estágio é categórica ordinal e sua normalização já está embutida na definição das distâncias, conforme descrito na Equação 3.7, por isso não é necessário normalizar estes valores. Isso resulta nos novos dados apresentados na Tabela 2.

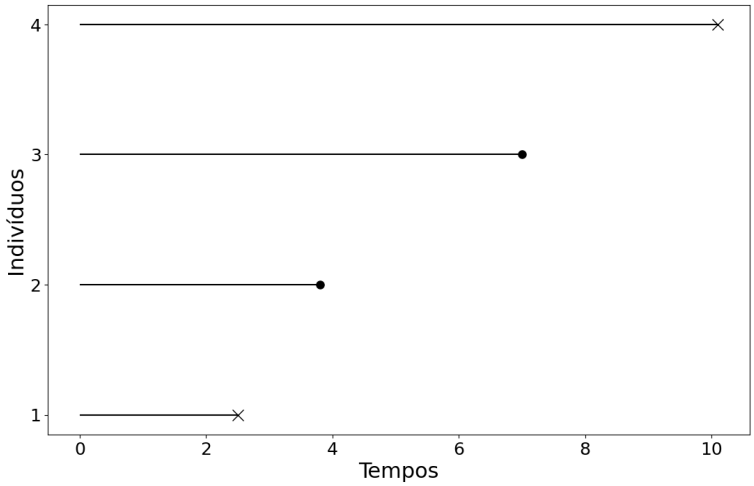
Tabela 2 – Dados fictícios normalizados

id	**Tempo**	**Falha**	**Estágio**	**Idade**
1	2.5	0	1	1
2	3.8	1	4	0
3	7.0	1	2	0.388
4	10.0	0	3	0.777
5	-	-	1	0.888

Fonte: Elaborada pela autora.

A padronização consiste em ajustar os dados para variarem em um intervalo específico, geralmente entre zero e um. Isso significa que o valor máximo absoluto de cada recurso é ajustado para 1 e o valor mínimo absoluto é ajustado para 0.

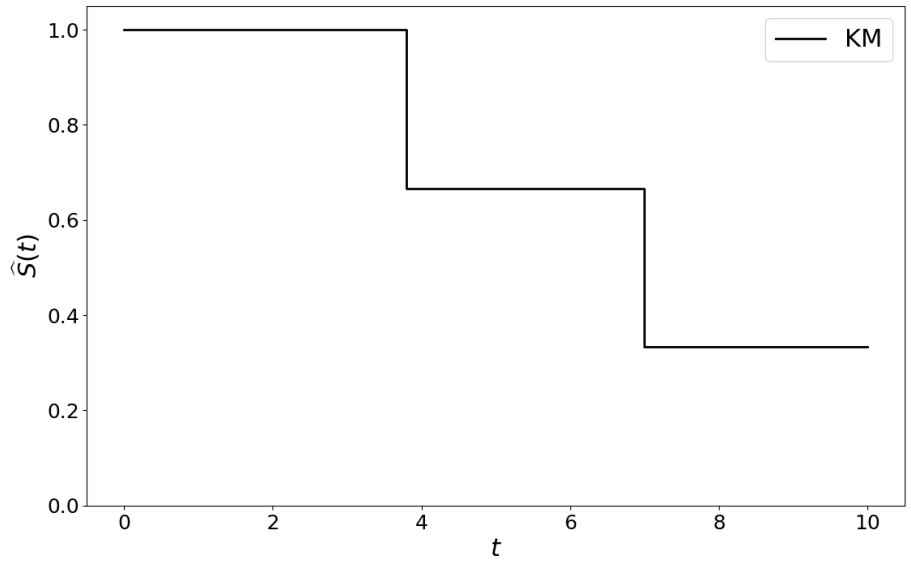
Figura 3 – Linha do tempo para os dados fictícios em que • representa falha e × representa censura



Fonte: Elaborada pela autora.

Por fim, podemos visualizar a linha do tempo representada na Figura 3, na qual observamos o tempo de falha ou censura de cada paciente. Adicionalmente, utilizando o estimador clássico de Kaplan-Meier (Equação 2.9), conseguimos traçar a curva de sobrevivência para esta pequena amostra de pacientes, conforme ilustrado na Figura 4. A partir da função de sobrevivência, podemos também obter o tempo médio, definido pela Equação 2.7, como $t_m = 6.93$.

Figura 4 – Estimativa de Kaplan-Meier para os dados fictícios elaborada utilizando a biblioteca *lifelines* disponível em Python (Davidson-Pilon, 2019)



Fonte: Elaborada pela autora.

Para o vetor de covariáveis $\mathbf{x}_5 = (1, 0.888)^\intercal$, é necessário estimar a função de sobrevivência para um paciente com estas características utilizando a Equação 4.1, em que $n = n' = 4$

$$\widehat{S}(t|\mathbf{x}_5) = \prod_{i=1}^{4} \left(1 - \frac{\sum_{j=1}^{4} s_w(\mathbf{x}_5, \mathbf{x}_j)\delta_j \mathbb{1}\{t_j = t_i\}}{\sum_{j=1}^{4} s_w(\mathbf{x}_5, \mathbf{x}_j)\mathbb{1}\{t_j \geq t_i\}}\right)^{\delta_i \mathbb{1}\{t_i \leq t\}} \tag{4.5}$$

$$= \left(1 - \frac{\sum_{j=1}^{4} s_w(\mathbf{x}_5, \mathbf{x}_j)\delta_j \mathbb{1}\{t_j = t_2\}}{\sum_{j=1}^{4} s_w(\mathbf{x}_5, \mathbf{x}_j)\mathbb{1}\{t_j \geq t_2\}}\right)^{\mathbb{1}\{t_2 \leq t\}} \left(1 - \frac{\sum_{j=1}^{4} s_w(\mathbf{x}_5, \mathbf{x}_j)\delta_j \mathbb{1}\{t_j = t_3\}}{\sum_{j=1}^{4} s_w(\mathbf{x}_5, \mathbf{x}_j)\mathbb{1}\{t_j \geq t_3\}}\right)^{\mathbb{1}\{t_3 \leq t\}}$$

$$= \left(1 - \frac{s_w(\mathbf{x}_5, \mathbf{x}_2)}{s_w(\mathbf{x}_5, \mathbf{x}_2) + s_w(\mathbf{x}_5, \mathbf{x}_3) + s_w(\mathbf{x}_5, \mathbf{x}_4)}\right)^{\mathbb{1}\{t_2 \leq t\}} \left(1 - \frac{s_w(\mathbf{x}_5, \mathbf{x}_3)}{s_w(\mathbf{x}_5, \mathbf{x}_3) + s_w(\mathbf{x}_5, \mathbf{x}_4)}\right)^{\mathbb{1}\{t_3 \leq t\}}.$$

Para este exemplo, a forma da função de similaridade será fixada na forma exponencial (EX), dada por:

$$s_w(\mathbf{x}_i, \mathbf{x}_j) = e^{-d_w(\mathbf{x}_i, \mathbf{x}_j)}, \tag{4.6}$$

em que d_w representa a distância total obtida a partir da Equação 3.7, a qual, a partir destes dados, pode ser calculada da seguinte maneira

$$\begin{aligned} d_w &= d_w^{CO} + d_w^{Q} \\ &= \frac{w_1(x_i^1 - x_j^1)^2}{(x_{max}^1 - x_{min}^1)^2} + w_2(x_i^2 - x_j^2)^2, \end{aligned} \tag{4.7}$$

em que w_1 é o peso associado à covariável Estágio e w_2 é o peso associado à variável Idade. Portanto, a função de similaridade fica definida em um vetor bidimensional de parâmetros.

Para estimar os valores dos parâmetros w, podemos utilizar o método de maximização da verossimilhança definido na Equação 4.2

$$L(w) = \widehat{S}(t_1|\mathbf{x}_1)\widehat{P}(t_2|\mathbf{x}_2)\widehat{P}(t_3|\mathbf{x}_3)\widehat{S}(t_4|\mathbf{x}_4). \tag{4.8}$$

Calculando os termos desta expressão separadamente, obtemos:

$\widehat{S}(t_1|\mathbf{x}_1) = 1$

$\widehat{P}(t_2|\mathbf{x}_2) = \widehat{S}(t_2^-|\mathbf{x}_2)\frac{\sum_{j=1}^{4} s_w(\mathbf{x}_2,\mathbf{x}_j)\delta_j \mathbb{1}\{t_j = t_2\}}{\sum_{j=1}^{4} s_w(\mathbf{x}_2,\mathbf{x}_j)\mathbb{1}\{t_j \geq t_2\}} = \frac{s_w(\mathbf{x}_2,\mathbf{x}_2)}{s_w(\mathbf{x}_2,\mathbf{x}_2)+s_w(\mathbf{x}_2,\mathbf{x}_3)+s_w(\mathbf{x}_2,\mathbf{x}_4)}$

$\widehat{P}(t_3|\mathbf{x}_3) = \widehat{S}(t_3^-|\mathbf{x}_3)\frac{\sum_{j=1}^{4} s_w(\mathbf{x}_3,\mathbf{x}_j)\delta_j \mathbb{1}\{t_j = t_3\}}{\sum_{j=1}^{4} s_w(\mathbf{x}_3,\mathbf{x}_j)\mathbb{1}\{t_j \geq t_3\}} = \left(1 - \frac{\sum_{j=1}^{4} s_w(\mathbf{x}_3,\mathbf{x}_j)\delta_j \mathbb{1}\{t_j = t_2\}}{\sum_{j=1}^{4} s_w(\mathbf{x}_3,\mathbf{x}_j)\mathbb{1}\{t_j \geq t_2\}}\right)^{\mathbb{1}\{t_2 \leq t_3^-\}} \frac{s_w(\mathbf{x}_3,\mathbf{x}_3)}{s_w(\mathbf{x}_3,\mathbf{x}_3)+s_w(\mathbf{x}_3,\mathbf{x}_4)}$

$= \left(1 - \frac{s_w(\mathbf{x}_3,\mathbf{x}_2)}{s_w(\mathbf{x}_3,\mathbf{x}_2)+s_w(\mathbf{x}_3,\mathbf{x}_3)+s_w(\mathbf{x}_3,\mathbf{x}_4)}\right) \frac{s_w(\mathbf{x}_3,\mathbf{x}_3)}{s_w(\mathbf{x}_3,\mathbf{x}_3)+s_w(\mathbf{x}_3,\mathbf{x}_4)}$

$$\widehat{S}(t_4|\mathbf{x}_4) = \left(1 - \frac{\sum_{j=1}^{4} s_w(\mathbf{x}_4,\mathbf{x}_j)\delta_j \mathbb{1}\{t_j=t_2\}}{\sum_{j=1}^{4} s_w(\mathbf{x}_4,\mathbf{x}_j)\mathbb{1}\{t_j\geq t_2\}}\right)^{\mathbb{1}\{t_2\leq t_4\}} \left(1 - \frac{\sum_{j=1}^{4} s_w(\mathbf{x}_4,\mathbf{x}_j)\delta_j \mathbb{1}\{t_j=t_3\}}{\sum_{j=1}^{4} s_w(\mathbf{x}_4,\mathbf{x}_j)\mathbb{1}\{t_j\geq t_3\}}\right)^{\mathbb{1}\{t_3\leq t_4\}} =$$
$$\left(1 - \frac{s_w(\mathbf{x}_4,\mathbf{x}_2)}{s_w(\mathbf{x}_4,\mathbf{x}_2)+s_w(\mathbf{x}_4,\mathbf{x}_3)+s_w(\mathbf{x}_4,\mathbf{x}_4)}\right)\left(1 - \frac{s_w(\mathbf{x}_4,\mathbf{x}_3)}{s_w(\mathbf{x}_4,\mathbf{x}_3)+s_w(\mathbf{x}_4,\mathbf{x}_4)}\right)$$

Neste ponto, torna-se necessário definir a matriz de similaridade empírica, em que cada termo pode ser calculado utilizando a Equação 4.6.

$$s_w = \begin{bmatrix} s_w(\mathbf{x}_1,\mathbf{x}_1) & s_w(\mathbf{x}_1,\mathbf{x}_2) & s_w(\mathbf{x}_1,\mathbf{x}_3) & s_w(\mathbf{x}_1,\mathbf{x}_4) & s_w(\mathbf{x}_1,\mathbf{x}_5) \\ s_w(\mathbf{x}_2,\mathbf{x}_1) & s_w(\mathbf{x}_2,\mathbf{x}_2) & s_w(\mathbf{x}_2,\mathbf{x}_3) & s_w(\mathbf{x}_2,\mathbf{x}_4) & s_w(\mathbf{x}_2,\mathbf{x}_5) \\ s_w(\mathbf{x}_3,\mathbf{x}_1) & s_w(\mathbf{x}_3,\mathbf{x}_2) & s_w(\mathbf{x}_3,\mathbf{x}_3) & s_w(\mathbf{x}_3,\mathbf{x}_4) & s_w(\mathbf{x}_3,\mathbf{x}_5) \\ s_w(\mathbf{x}_4,\mathbf{x}_1) & s_w(\mathbf{x}_4,\mathbf{x}_2) & s_w(\mathbf{x}_4,\mathbf{x}_3) & s_w(\mathbf{x}_4,\mathbf{x}_4) & s_w(\mathbf{x}_4,\mathbf{x}_5) \\ s_w(\mathbf{x}_5,\mathbf{x}_1) & s_w(\mathbf{x}_5,\mathbf{x}_2) & s_w(\mathbf{x}_5,\mathbf{x}_3) & s_w(\mathbf{x}_5,\mathbf{x}_4) & s_w(\mathbf{x}_5,\mathbf{x}_5) \end{bmatrix} \tag{4.9}$$

$$= \begin{bmatrix} 1 & e^{-(w_1+w_2)} & e^{-(0.111w_1+0.374w_2)} & e^{-(0.444w_1+0.049w_2)} & e^{-0.012w_2} \\ e^{-(w_1+w_2)} & 1 & e^{-(0.444w_1+0.150w_2)} & e^{-(0.111w_1+0.604w_2)} & e^{-(w_1+0.790w_2)} \\ e^{-(0.111w_1+0.374w_2)} & e^{-(0.444w_1+0.150w_2)} & 1 & e^{-(0.111w_1+0.151w_2)} & e^{-(0.111w_1+0.250w_2)} \\ e^{-(0.444w_1+0.049w_2)} & e^{-(0.111w_1+0.604w_2)} & e^{-(0.111w_1+0.151w_2)} & 1 & e^{-(0.444w_1+0.012w_2)} \\ e^{-0.012w_2} & e^{-(w_1+0.790w_2)} & e^{-(0.111w_1+0.250w_2)} & e^{-(0.444w_1+0.012w_2)} & 1 \end{bmatrix}.$$

Nota-se que a matriz é simétrica e sua diagonal é igual a 1, uma vez que cada vetor de covariáveis é idêntico a si.

Substituindo os valores correspondentes da matriz, obtemos a expressão final para a verossimilhança:

$$L(w) = \frac{1}{1+e^{-(0.444w_1+0.150w_2)}+e^{-(0.111w_1+0.604w_2)}} \tag{4.10}$$
$$\left(1 - \frac{e^{-(0.444w_1+0.150w_2)}}{e^{-(0.444w_1+0.150w_2)}+1+e^{-(0.111w_1+0.151w_2)}}\right)\frac{1}{1+e^{-(0.111w_1+0.151w_2)}}$$
$$\left(1 - \frac{e^{-(0.111w_1+0.604w_2)}}{e^{-(0.111w_1+0.604w_2)}+e^{-(0.111w_1+0.151w_2)}+1}\right)\left(1 - \frac{e^{-(0.111w_1+0.151w_2)}}{e^{-(0.111w_1+0.151w_2)}+1}\right).$$

É possível calcular o logaritmo natural dessa expressão para simplificar alguns termos e converter as funções exponenciais em formas mais fáceis de manipular. Ao fazer isso, obtemos:

$$\begin{aligned} \ln[L(w)] &= -\ln[1+e^{-(0.444w_1+0.150w_2)}+e^{-(0.111w_1+0.604w_2)}] \\ &- \ln[e^{-(0.444w_1+0.150w_2)}+1+e^{-(0.111w_1+0.151w_2)}] \\ &- \ln[e^{-(0.111w_1+0.604w_2)}+e^{-(0.111w_1+0.151w_2)}+1]. \end{aligned} \tag{4.11}$$

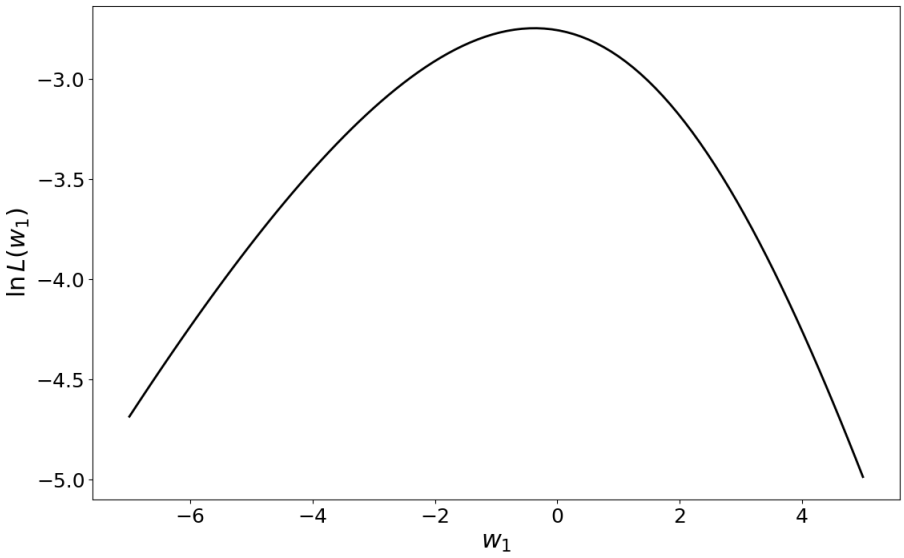
Ao analisar a Equação 4.11, concluímos que sempre que aumentamos w_1 e w_2, a função aumenta, indicando que a função de verossimilhança não possui um máximo. Portanto, para o nosso método de estimação, torna-se necessário incluir uma nova restrição aos pesos. Assumimos, então, que $w_1 + w_2 = 1$, ou seja, iremos considerar o peso relativo das covariáveis e não os pesos absolutos.

Desse modo, podemos reescrever a Equação 4.11 considerando que $w_2 = 1 - w_1$,

$$\begin{aligned} \ln[L(w)] &= -\ln[1 + e^{-(0.294w_1+0.150)} + e^{+(0.494w_1-0.604)}] \\ &- \ln[e^{-(0.294w_1+0.150)} + 1 + e^{+(0.040w_1-0.151)}] \\ &- \ln[e^{+(0.494w_1-0.604)} + e^{+(0.040w_1-0.151)} + 1]. \end{aligned} \tag{4.12}$$

A função pode ser visualizada na Figura 5. Observa-se que o valor de w_1 que maximiza a função de verossimilhança é negativo. No entanto, devido às restrições estabelecidas anteriormente, só podemos avaliar os valores de w_1 que estão dentro do intervalo de $[0, 1]$.

Figura 5 – Representação gráfica do logaritmo da função de verossimilhança

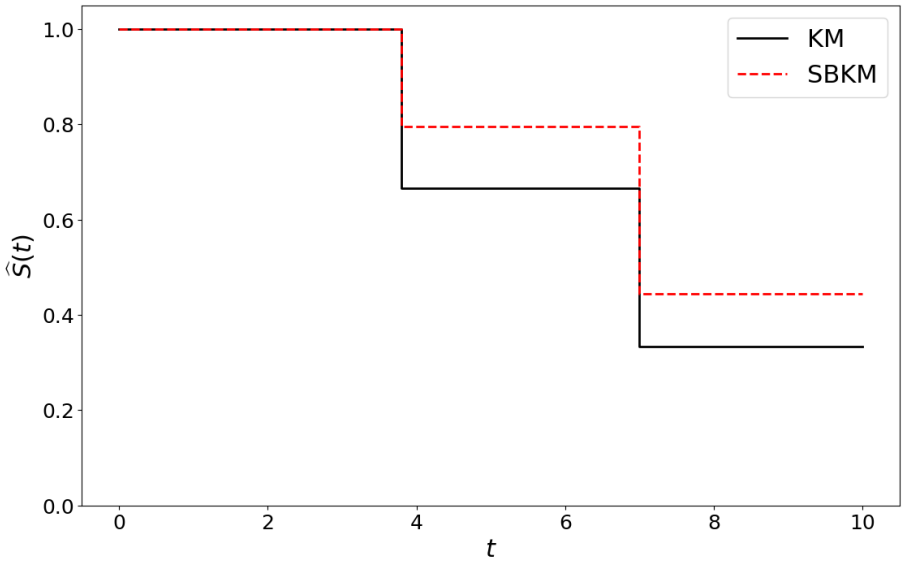


Fonte: Elaborada pela autora.

Assim, determinamos $w_1 = 0$ e $w_2 = 1$. Em seguida, podemos calcular a função de sobrevivência $\widehat{S}(t|\mathbf{x}_5)$ utilizando a Equação 4.5, representar graficamente a curva correspondente na Figura 6, e calcular o tempo médio para este indivíduo como $t_m = 7.68$.

$$\widehat{S}(t|\mathbf{x}_5) = (0.796)^{\mathbb{1}\{t_2 \leq t\}}(0.556)^{\mathbb{1}\{t_3 \leq t\}} \tag{4.13}$$

Figura 6 – Curva de sobrevivência obtida para o 5º paciente pelo estimador SBKM em comparação com a estimativa de KM



Fonte: Elaborada pela autora.

5 MÉTODOS E MATERIAIS

5.1 Dados de Sobrevivência

Há diversos exemplos de aplicação dos modelos de análise de sobrevivência. Na esfera médica, eles são amplamente empregados na identificação de fatores prognósticos para doenças, assim como na comparação de tratamentos. Além disso, na indústria, existem diversas aplicações da análise de sobrevivência. Por exemplo, a técnica de manutenção preditiva consiste em prever o momento da falha de um equipamento, permitindo à equipe de manutenção antecipar-se e evitar possíveis falhas.

Para ilustrar os métodos abordados neste trabalho, utilizaremos dois conjuntos de dados reais. A primeira base de dados está relacionada a uma aplicação no mercado financeiro, mais especificamente no setor de empréstimo pessoal. O risco de crédito, que se refere à probabilidade de um mutuário não conseguir reembolsar um empréstimo, é uma preocupação significativa para instituições financeiras. Ao longo dos anos, essas instituições desenvolveram diversas formas de quantificar esse risco, visando limitar sua exposição.

A análise de sobrevivência, neste contexto, vai além da simples modelagem de se um mutuário pagará ou não o empréstimo, permitindo determinar quando esse pagamento ocorrerá. Considerando o reembolso total do empréstimo como um evento explícito, a situação em que o empréstimo ainda não foi pago pode ser definida como um evento censurado.

Para exemplificar, faremos uso do conjunto de dados German Credit (CREDIT), originalmente fornecido pelo Professor Dr. Hans Hofmann da Universidade de Hamburgo e disponível no Repositório de Aprendizado de Máquina da UCI (Hofmann, 1994). Este conjunto de dados compreende informações pessoais e sociodemográficas de diversos mutuários, além da duração em meses desde a data do empréstimo e o status de reembolso total. Todos os detalhes disponíveis nesses dados estão listados na Tabela 3.

Tabela 3 – Variáveis utilizadas do conjunto de dados CREDIT

Variável	**Tipo**	**Descrição**
duration	Numérico	Duração em meses
full_repaid	Categórico	Especifica se o empréstimo foi totalmente reembolsado
age	Numérico	Idade do mutuário (em anos)
foreign_worker	Categórico	Indica se o mutuário é um trabalhador estrangeiro
personal_status	Categórico	Gênero e estado civil
people_liable	Numérico	Número de pessoas responsáveis pela manutenção
telephone	Categórico	Indica se o mutuário possui um telefone
employment_years	Categórico	Anos (em intervalos) de trabalho no emprego atual
job	Categórico	Situação de emprego
housing	Categórico	Situação residencial do mutuário
present_residence	Numérico	Anos morando na residência atual
amount	Numérico	Quantidade de dinheiro emprestado
installment_rate	Numérico	Porcentagem do valor emprestado que será cobrado por um credor de um devedor
purpose	Categórico	Motivo para obter um empréstimo
checking_account_status	Categórico	Situação da conta-corrente
credit_history	Categórico	Histórico de crédito do mutuário
number_of_credits	Numérico	Número de créditos existentes neste banco
savings_account_status	Categórico	Situação da conta poupança
property	Categórico	Tipo de ativos valiosos que o mutuário possui

Fonte: Elaborada pela autora.

A segunda base de dados utilizada neste trabalho pertence ao contexto mais tradicional da área médica. Ele se concentra na ocorrência de óbito em pacientes e foi adquirido a partir de um estudo clínico conduzido por Knaus *et al.* (1995). O objetivo principal desse estudo era desenvolver e validar um modelo prognóstico capaz de estimar a sobrevivência de adultos doentes ao longo de um período de 180 dias, enquanto estavam hospitalizados. Dessa forma, o objetivo central é prever o tempo decorrido até a morte de cada paciente. Este experimento foi conduzido como parte do projeto "Study to Understand Prognoses and Preferences for Outcomes and Risks of Treatments" (SUPPORT). Os dados abrangem diversas variáveis, incluindo diagnóstico, idade, presença de câncer, função neurológica e medidas fisiológicas registradas no terceiro dia após a admissão no estudo. Banco de descrição das covariáveis estão disponíveis na Tabela 4.

Tabela 4: Variáveis utilizadas do conjunto de dados SUPPORT

Variável	**Tipo**	**Descrição**
d_time	Numérico	Duração em meses
death	Categórico	Indica se o paciente veio a óbito
age	Numérico	Idade do paciente
sex	Categórico	Gênero do paciente
race	Categórico	Raça do paciente
num_co	Numérico	Número de comorbidades
diabetes	Categórico	Indica se o paciente tem diabetes
dementia	Categórico	Indica se o paciente tem demência
ca	Categórico	Estado de câncer
meanbp	Numérico	Pressão arterial média
hrt	Numérico	Frequência cardíaca
resp	Numérico	Taxa respiratória
temp	Numérico	Temperatura corporal do paciente em graus Celsius
wblc	Numérico	Contagem de glóbulos brancos no sangue
sod	Numérico	Concentração de sódio sérico
crea	Numérico	Concentração de creatinina sérica

Fonte: Elaborada pela autora.

Em ambas as bases, realizamos o processamento dos dados. Para as covariáveis

numéricas, reescalamos os valores para a escala de 0 a 1. Quanto às covariáveis categóricas ordinais, codificamos para uma escala numérica, mantendo a ordem clara das categorias correspondentes. Para o estimador proposto SBKM, não é necessário nenhum tratamento adicional das covariáveis categóricas nominais. No entanto, para os demais modelos utilizados como referência, abordados posteriormente neste capítulo, foi necessário realizar a codificação em covariáveis indicadoras do tipo *dummy*, que são variáveis binárias.

Em suma, ao explorar conjuntos de dados provenientes de contextos bastante diferentes, tanto no âmbito financeiro quanto clínico, nosso propósito foi examinar como os estimadores e modelos se comportam em dados com diferentes características, como podemos visualizar na Tabela 5. Também visamos ilustrar a amplitude de áreas e temas nos quais podemos aproveitar as técnicas de análise de sobrevivência, destacando a versatilidade e aplicabilidade dessas abordagens.

Tabela 5: Descrição das características dos conjuntos de dados após o pré-processamento

Dados	**Número de amostras**	**Número de covariáveis**	**Taxa de censura**
CREDIT	1000	17	30.0%
SUPPORT	8873	14	31.97%

Fonte: Elaborada pela autora.

5.2 Metodologia

Nossos experimentos foram projetados visando aprofundar a compreensão do desempenho prático dos métodos de análise de sobrevivência, especificamente no que diz respeito à precisão das previsões em distintos conjuntos de dados reais. Para conduzir essa análise, adotamos uma estratégia convencional de aprendizado de máquina, dividindo a base de dados em conjuntos de treinamento e teste (Xu; Goodacre, 2018). Na fase de treinamento, obtemos as estimativas dos parâmetros, ou seja, o vetor de pesos $\mathbf{w}$, e usamos estas estimativas na base de teste para gerar as previsões necessárias. A estimação dos parâmetros foi feita com o auxílio da biblioteca *SciPy* do Python, empregando o método de otimização "SLSQP" (Programação Sequencial de Mínimos Quadrados) (Kraft, 1988).

Para ambos os conjuntos de dados, implementamos uma divisão aleatória de treinamento/teste, sendo de 70%/30%, e reservamos 20% dos dados de treinamento para a fase de validação. A introdução dessa etapa de validação possibilita a avaliação e otimização do desempenho do modelo antes da análise definitiva na base de teste. Neste trabalho, a base de

validação foi empregada para a seleção das covariáveis mais relevantes, bem como para determinar os hiperparâmetros mais eficazes entre os diferentes modelos comparativos utilizados. Em outras palavras, a base de validação é empregada para definir os hiperparâmetros e também para monitorar o fenômeno de sobreajuste, em inglês *overfitting*, onde o modelo se ajusta muito bem aos dados de treinamento, mas falha em generalizar para novos dados. O *overfitting* pode ocorrer quando uma função está muito alinhada a um conjunto limitado de unidades amostrais, tornando o modelo útil apenas para os dados específicos do conjunto de treinamento. Por fim, a base de teste é destinada a avaliar o potencial de generalização do estimador para outros conjuntos de dados além do conjunto de treinamento original.

Para a seleção das covariáveis, utilizamos o algoritmo *Stepwise Forward Selection*, amplamente empregado em modelos de regressão. Este algoritmo seleciona gradualmente as variáveis mais significativas ou relevantes para o modelo (Zhang, 2016). O método constrói o modelo adicionando variáveis uma a uma, começando pela que proporciona o maior ganho em termos de ajuste. A estatística utilizada no algoritmo foi o Concordance Index (CI) na base de validação, com um critério de parada estabelecido em uma melhoria mínima de 0.01 nessa métrica.

Para avaliar a recuperação dos parâmetros obtidos, das métricas de desempenho e outras medidas, realizamos a construção de um intervalo de confiança de 95% usando o método de *Bootstrap* (Wehrens *et al.*, 2000). Isso envolveu a reamostragem da base de teste 200 vezes com reposição. Para cada amostra obtida, calculamos as métricas de interesse e fornecemos os percentis 2.5 e 97.5 como indicadores dos limites do intervalo de confiança. Esse procedimento oferece uma abordagem robusta e eficaz para avaliar a variabilidade das métricas e fornecer estimativas mais confiáveis do desempenho dos modelos. Escolhemos critérios semelhantes aos utilizados na literatura (Chen, 2024).

5.3 Modelos de Sobrevivência de Referência

Tradicionalmente, na literatura, foram amplamente investigadas diversas abordagens estatísticas para lidar com o desafio dos dados censurados. Além disso, vários algoritmos de aprendizado de máquina também foram adaptados para efetivamente tratar dados de sobrevivência e abordar outras questões complexas comumente encontradas em conjuntos de dados reais. Para avaliar a eficácia do estimador proposto, selecionamos diversos modelos de referência, tanto estatísticos quanto de aprendizado de máquina, para comparação dos resultados. A

escolha dos modelos de referência foi orientada pela facilidade de implementação, utilizando as bibliotecas *lifelines* (Davidson-Pilon, 2019) e *scikit-survival* (Pölsterl, 2020) disponíveis na linguagem Python, bem como pelo artigo de revisão publicado por Wang *et al.* (2019). Os modelos selecionados foram os seguintes:

- Cox Proportional-Hazards (COX):

  O modelo de Cox é o modelo de sobrevivência semi-paramétrico mais amplamente empregado na literatura (Cox, 1972). Esse modelo assume que o risco relativo de um evento, como a morte, permanece constante ao longo do tempo, caracterizando-o como um modelo de riscos proporcionais. Logo, a taxa de risco é tratada como uma constante, e todos os sujeitos compartilham a mesma função de risco de linha de base.

- Elastic-Net Cox (EN-COX):

  Uma extensão do modelo de Cox, proposta por Simon *et al.* (2011), que combina as características de regularização do Lasso e Ridge com o potencial para realizar a seleção de covariáveis e lidar, simultaneamente, com a correlação entre elas.

- Distribuição de Weibull (WEIBULL):

  O modelo de regressão Weibull é uma das formas mais populares de modelo de regressão paramétrica. A distribuição Weibull é particularmente popular na análise de sobrevivência, por poder modelar com precisão o tempo até a falha de eventos reais e é suficientemente flexível, apesar de ter apenas dois parâmetros.

- Survival Tree (ST):

  As árvores de sobrevivência são uma forma de árvores de classificação e regressão adaptadas para lidar com dados censurados. A intuição básica por trás dos modelos de árvore é particionar recursivamente os dados com base em um determinado critério de divisão, e os objetos que são semelhantes entre si com base no evento de interesse serão colocados no mesmo nó. A criação de árvores de sobrevivência foi discutida e publicada pela primeira vez por Gordon; Olshen (1985).

- Random Survival Forest (RSF):

  A floresta aleatória é um método de tipo *ensemble*, que consiste na construção de vários modelos de aprendizado de máquina, onde o resultado de cada modelo é utilizado na definição de um único resultado, resultando em uma previsão final única e com melhor desempenho. Especificamente propostas para fazer previsões utilizando modelos estruturados em árvore, as Florestas Aleatórias são criadas gerando múltiplas árvores de decisão com

amostras bootstrap dos dados, e selecionando aleatoriamente subconjuntos de variáveis em cada divisão (Breiman, 2001). Por sua vez, a Floresta de Sobrevivência Aleatória (RSF) estende o método da Floresta Aleatória de Breiman, utilizando uma floresta de árvores de sobrevivência para previsão (Ishwaran *et al.*, 2008).

- Gradient-Boosted Cox (GB-COX):

  O algoritmo *Gradient Boosting* é uma técnica integrada ao conjunto de métodos *ensemble*, amplamente utilizada para combinar preditores fracos em uma soma ponderada representando um preditor forte. Os resíduos são definidos com base neste algoritmo e ajustados iterativamente. Hothorn *et al.* (2006) estende este método para minimizar a função de risco ponderada, aplicando-o ao ajuste de um modelo de riscos proporcionais de Cox. Aqui, os resíduos são iterativamente ajustados para aprimorar a precisão ao longo do tempo.

Detalhes sobre a otimização dos hiperparâmetros de cada um dos modelos utilizados poderão ser encontrados descritos no Apêndice A desta dissertação.

5.4 Métricas de Avaliação

Usualmente, os modelos preditivos em análise de sobrevivência mapeiam o conjunto de covariáveis $\mathbf{x}_i$ para um score de risco de um determinado indíviduo experienciar o evento $\eta_i \in \mathbb{R}$, assim como para sua função de sobrevivência $S(t|\mathbf{x}_i)$. Especificamente, lidamos com pontuações de "risco" que nada mais são que medidas da probabilidade de que um evento aconteça a um indivíduo.

Em particular, um "evento", ou falha, é uma ocorrência inevitável que altera permanentemente o status de um indivíduo. Por outro lado, se paramos de observar um indivíduo antes que ele experimente um evento, teremos apenas uma observação parcial chamada de censura. Recordamos que, no conjunto de dados, são registrados tempos de censura e tempos de falha e fazemos a distinção entre os dois utilizando a variável binária δ_i, que é igual a 1 se e somente se o i-ésimo sujeito experimentou o evento em algum momento do tempo t_i, e igual a 0 se o mesmo foi censurado antes disso. Dessa forma, só podemos dizer que o evento acontece após a censura, mas o tempo exato é desconhecido.

Neste cenário, avaliar a qualidade do ajuste do modelo torna-se mais difícil, pois a variável alvo não é totalmente observada devido à presença de censuras nos dados, e métricas de regressão tradicionais como erro quadrático médio, erro médio absoluto e erro percentual médio absoluto podem ser inadequadas.

Dito isso, nesta sessão, serão apresentadas as métricas de avaliação mais utilizadas no contexto de sobrevivência para medir diferentes aspectos do desempenho de um determinado modelo. Estas métricas incorporam tanto eventos observados quanto casos censurados.

5.4.1 Índice de Concordância

A métrica de avaliação de modelos de sobrevivência mais utilizada é o índice de concordância, do inglês CI, também conhecido como *C-index*. Esta medida avalia quão bem um modelo é capaz de ordenar as instâncias de dados corretamente conforme os tempos de sobrevivência e com base nas pontuações de risco individuais.

O CI é definido como a proporção de todos os pares comparáveis nos quais as previsões e os resultados são concordantes. Duas amostras i e j são comparáveis se a amostra com menor tempo observado sofreu o evento, ou seja, se $t_j > t_i$ e $\delta_i = 1$, onde δ_i é o indicador de evento binário. Suponha que valores mais altos de η implicam um valor menor para t, então um par comparável é concordante se o risco estimado por um modelo de sobrevivência é maior para indivíduos com menor tempo de sobrevivência, isto é, $\eta_i > \eta_j$ e $t_i < t_j$, caso contrário, o par é discordante (Alabdallah *et al.*, 2024).

Desse modo, podemos considerar o *C-index* como a seguinte probabilidade, condicionada à ordem relativa dos eventos (Longato *et al.*, 2020).

$$CI = P(\eta_i > \eta_j | t_i < t_j) \tag{5.1}$$

Para um modelo perfeitamente discriminativo, se você escolher dois sujeitos comparáveis aleatoriamente (η_i, t_i) e (η_j, t_j), então, aquele com o maior valor de η terá, com probabilidade 1, um tempo de sobrevivência inferior. Logo, o índice de concordância é uma medida de quão discriminante é um modelo, onde discriminação significa que o modelo consegue prever, com alta confiabilidade, entre duas instâncias qual delas terá o menor tempo de sobrevivência (Uno *et al.*, 2011).

Podemos calcular essa probabilidade computacionalmente a partir da seguinte fórmula:

$$\widehat{CI} = \frac{\sum_{i,j} \mathbb{1}\{t_i < t_j\} \mathbb{1}\{\eta_i > \eta_j\} \delta_i}{\sum_{i,j} \mathbb{1}\{t_i < t_j\} \delta_i}, \tag{5.2}$$

onde:

- η_i é o score de risco do indivíduo i

- $\mathbb{1}\{t_i < t_j\} = 1$ se $t_i < t_j$ c.c. 0
- $\mathbb{1}\{\eta_i > \eta_j\} = 1$ se $\eta_i > \eta_j$ c.c. 0

Consequentemente, $\widehat{CI} = 1$ corresponde à melhor previsão do modelo e o índice $\widehat{CI} = 0.5$ representa uma previsão aleatória.

O SBKM, proposto neste trabalho na Equação 4.1, tem a capacidade de estimar a função de sobrevivência condicional. A partir dessa estimativa, torna-se viável calcular o tempo médio t_m e o tempo mediano $t_{0.5}$ para cada indivíduo da amostra. O tempo de falha estimado para cada indivíduo, seja por meio da média ou da mediana, é então empregado como o indicador de risco η para o cálculo do CI.

Quantificando a correlação entre as previsões de risco e os tempos dos eventos, entendemos que maximizar esta métrica significa melhorar a qualidade da discriminação entre eventos precoces, associados a contextos de maior risco, e ocorrências posteriores (Longato *et al.*, 2020). A sua capacidade de resumir três dimensões diferentes de previsões (risco, ocorrência de eventos e tempo) em um único número permite distinguir entre modelos "bem comportados" e modelos "praticamente aleatórios" e, por esse motivo, é a métrica mais amplamente utilizada para a avaliação global de modelos em análise de sobrevivência.

Entretanto, o índice de concordância avalia a ordenação relativa dos tempos dos eventos dos indivíduos. Portanto, é invariante à escala e ao deslocamento (ou seja, pode-se multiplicar por uma constante positiva ou adicionar uma constante e o ranking não mudará). Um modelo maximizado para o índice de concordância não fornece necessariamente bons tempos previstos, mas fornecerá boas ordenações previstas.

5.4.2 Brier Score

O Brier Score (BS) é uma extensão do erro quadrático médio para dados censurados à direita e é principalmente usado para avaliar a precisão de uma função de sobrevivência prevista em um determinado momento t. Desse modo, temos uma representação das distâncias quadradas médias entre o estado de sobrevivência observado e a probabilidade de sobrevivência prevista.

O *C-index* é provavelmente a medida de discriminação mais comum. No entanto, o índice de concordância ignora os valores reais das pontuações de risco previstas por ser uma métrica de ranqueamento, e não é capaz de nos dizer nada sobre a calibração. Por outro lado, o *Brier Score* não é uma medida apenas do desempenho de discriminação ou de calibração, mas uma medida de desempenho geral, que incorpora ambos os aspectos de discriminação e

calibração de um modelo (Park *et al.*, 2021). Um modelo é dito bem calibrado quando suas previsões de probabilidade correspondem bem às frequências observadas.

Dado um conjunto de dados com n amostras, $\forall i \in [1,n]$, $(t_i, \delta_i, \mathbf{x}_i)$ é o formato de cada unidade amostral e a função de sobrevivência prevista é $\widehat{S}(t|\mathbf{x}_i)$. Na ausência de censuras, o BS pode ser calculada de forma que:

$$BS(t) = \frac{1}{n}\sum_{i=1}^{n}\left(\mathbb{1}\{t_i > t\} - \widehat{S}(t|\mathbf{x}_i)\right)^2. \tag{5.3}$$

No entanto, se o conjunto de dados contém amostras censuradas à direita, então é necessário ajustar a pontuação ponderando as distâncias quadradas usando o método de probabilidade inversa de censura ponderada, onde cada indivíduo i é ponderado pelo inverso de uma estimativa da probabilidade condicional de ter permanecido sem censura até o momento t (Kvamme; Borgan, 2023).

Isto significa que devemos estimar a função de sobrevivência para os tempos censurados. Considere $G(t) = P(C > t)$ a função de sobrevivência condicional dos tempos de censura calculados pelo método de Kaplan-Meier, onde C é o tempo de censura, mas agora consideramos os eventos de falha como observações "censuradas" e as observações censuradas como "falhas" (Satten; Datta, 2001). Neste caso, a estimativa de Kaplan-Meier de $G(t)$ será dada por

$$\widehat{G}(t) = \prod_{i=1}^{n}\left(1 - \frac{e_i}{g_i}\right)^{(1-\delta_i)\mathbb{1}\{t_i \leq t\}}, \tag{5.4}$$

em que $t_1 < t_2 ... < t_n$ são os n tempos ordenados, $e_i = \sum_{j=1}^{n}(1-\delta_j)\mathbb{1}\{t_j = t_i\}$ é o número de censuras em t_i, $j = 1, ..., n$ e $g_i = \sum_{j=1}^{n}\mathbb{1}\{t_j \geq t_i\}$ é o número de indivíduos sob risco em t_i, ou seja, os indivíduos que não falharam e não foram censurados até o instante imediatamente anterior a t_i.

Assim, temos que $1/\widehat{G}(t)$ representa a probabilidade inversa de censura utilizada para melhorar a precisão das estimativas em análises de dados censurados (Willems, 2014). Assim, podemos estimar BS dado por:

$$BS(t) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\left(0 - \widehat{S}(t|\mathbf{x}_i)\right)^2 \mathbb{1}\{t_i \leq t\}\delta_i}{\widehat{G}(t_i)} + \frac{\left(1 - \widehat{S}(t|\mathbf{x}_i)\right)^2 \mathbb{1}\{t_i > t\}}{\widehat{G}(t)}\right) \tag{5.5}$$

Neste ponto, o cálculo é feito separadamente para cada instante de tempo. A integração matemática de vários valores de BS obtidos em todos os instantes de tempo, chamada de Integrated Brier Score (IBS), pode então ser calculada como uma medida de desempenho

médio geral para o modelo de previsão para todos os momentos disponíveis $t_1 \leq t \leq t_{max}$ (Pölsterl, 2020). O IBS ao longo do intervalo $[t_1, t_{max}]$ pode ser definido como:

$$IBS = \int_{t_1}^{t_{max}} BS(t) d\omega(t), \tag{5.6}$$

onde a função de ponderação é $\omega(t) = t/t_{max}$.

Portanto, uma pontuação mais próxima de 0 indica melhor desempenho preditivo. Pensando em melhores práticas, um modelo útil terá um BS abaixo de 0.25 (Fotso *et al.*, 2019).

O BS é frequentemente usado para avaliar a calibração, pois se um modelo prevê um risco de 10% de ocorrer um evento em determinado momento, a frequência observada nos dados deve corresponder a esta porcentagem para um modelo bem calibrado. Além disso, o BS também é uma medida de discriminação: se um modelo tem o poder de prever pontuações de risco que nos permitam classificar corretamente a ordem dos eventos.

6 RESULTADOS

No Capítulo 5, apresentamos dois conjuntos de dados reais selecionados para aprofundar nossa compreensão sobre o comportamento do estimador de Kaplan-Meier baseado em Similaridade (SBKM) em diferentes cenários. Nosso objetivo é avaliar a eficácia do SBKM em comparação com outros modelos e algoritmos amplamente reconhecidos na análise de sobrevivência, verificando se o SBKM é competitivo em termos de qualidade e consistência das previsões. Para isso, investigamos diferentes configurações iniciais, variando a função de similaridade, as técnicas de normalização, as métricas de distância, bem como o tamanho da amostra e a taxa de censura, aspectos relacionados à própria base de dados. Esse processo é importante porque o SBKM possui diversos hiperparâmetros que precisam ser definidos previamente e podem influenciar significativamente seu desempenho. Ao testar essas configurações, é possível entender como cada hiperparâmetro afeta os resultados e ajustar o modelo conforme necessário, de modo a produzir melhores previsões em diferentes contextos de análise de sobrevivência. Para garantir que os resultados são representativos, calculamos intervalos de confiança com um nível de confiança de 95% para algumas das medidas reportadas, utilizando a técnica de *Bootstrap* descrita no Capítulo 5.

6.1 Base de dados CREDIT

Seguindo as etapas delineadas no capítulo anterior, a base de dados de crédito foi dividida aleatoriamente em conjuntos de treino, validação e teste, contendo 560, 140 e 300 amostras, respectivamente. Essa divisão de amostras foi fixada para garantir que todos os modelos fossem avaliados na mesma base de dados. A partir dos dados de treino selecionados, aplicamos o algoritmo *Stepwise Forward Selection*, utilizando o método de seleção baseado em nosso estimador SBKM para identificar as covariáveis mais relevantes para otimização do Índice de Concordância (CI) na base de validação, com o critério de parada estabelecido como uma melhoria significativa de pelo menos 0.01 nesta métrica. Após essa etapa, foram finalmente selecionadas duas covariáveis mais relevantes para as previsões, sendo elas “amount” e “installment_rate”, ambas variáveis numéricas.

Associado às covariáveis selecionadas, atribuímos o peso w_1 à covariável “amount” e o peso w_2 à covariável “installment_rate”. A partir deste ponto, podemos iniciar a exploração do SBKM a partir de diferentes configurações.

6.1.1 Normalização dos Pesos

No Capítulo 4, destacamos que nossa função de verossimilhança cresce indefinidamente com os pesos e não possui um valor máximo. Por essa razão, é necessário definir inicialmente um valor de normalização para os pesos. Embora a escolha óbvia seja $\sum_i w_i = 1$, como demonstrado no exemplo numérico, essa escolha pode não ser necessariamente a mais adequada, como será analisado adiante.

Outro aspecto relevante é que estamos interessados em obter estimativas com maior variabilidade para melhor distinguir os tempos de sobrevivência dos diversos indivíduos da base. Isso se deve à nossa expectativa de que o SBKM seja sensível às variações nas covariáveis, gerando estimativas distintas para indivíduos distintos. Em outras palavras, não buscamos um método que estime curvas de sobrevivência semelhantes para indivíduos que não compartilham características semelhantes.

Uma maneira indireta de avaliar a variabilidade das previsões é através do tempo de sobrevivência, ou tempo de falha, estimado para cada indivíduo na amostra de teste. Calculamos o tempo de falha previsto a partir das curvas de sobrevivência estimadas para cada indivíduo. Uma abordagem comum na literatura é utilizar o valor esperado, ou tempo médio t_m, conforme descrito na Equação 2.7. Neste trabalho, mantemos t_m como a estimativa padrão para o tempo de sobrevivência dos indivíduos, a menos que especificado de outra forma.

Dessa forma, empregamos uma métrica de dispersão para quantificar a variabilidade dos tempos de sobrevivência estimados. Escolhemos calcular o desvio padrão, cujo valor mínimo é 0, indicando ausência de variabilidade, ou seja, quando todos os valores são iguais à média. Portanto, computamos o desvio padrão dos tempos de falha estimados $\sigma(t_m)$ em cada amostra *bootstrap* e fornecemos a média acompanhada dos intervalos de confiança correspondentes.

Na Tabela 6, investigamos a normalização dos pesos. Analisamos como diferentes critérios de normalização influenciam os pesos estimados e, por conseguinte, sua ordem de importância, o que possibilita a identificação da covariável mais relevante para a estimação.

Tabela 6: Pesos estimados w_1, w_2 e desvio padrão dos tempos de falha estimados $\sigma(t_m)$ para diferentes condições de normalização $\sum_i w_i$ e funções de similaridade s_w

$\sum_i w_i$	s_w	w_1	w_2	$\sigma(t_m)$
1	EX	0.28	0.72	0.16 (0.14, 0.18)
	FR	0.38	0.62	0.17 (0.15, 0.19)
10	EX	6.01	3.99	2.72 (2.15, 3.32)
	FR	5.64	4.36	1.20 (1.04, 1.34)
100	EX	75.65	24.35	7.60 (6.69, 8.47)
	FR	65.09	34.91	3.43 (3.06, 3.73)
1000	EX	953.83	46.17	9.43 (8.30, 10.33)
	FR	763.58	236.42	6.05 (5.53, 6.53)

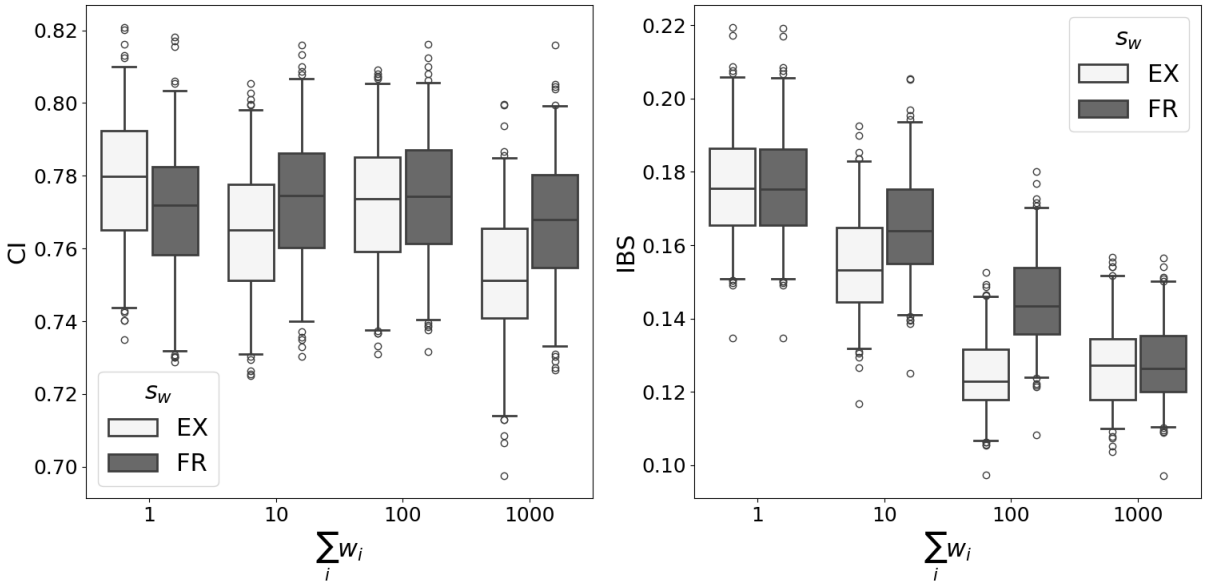
Fonte: Elaborada pela autora.

Nota-se que os pesos estimados, com exceção da condição de normalização $\sum_i w_i = 1$, seguem a mesma relação de importância, onde o valor de w_1 é sempre superior a w_2. Outro ponto importante é que a função de similaridade EX parece gerar estimativas mais variáveis em quase todos os cenários, excerto em $\sum_i w_i = 1$, onde não há diferença significativa.

Por último, observa-se que a normalização dos pesos parece influenciar a variabilidade das previsões, pois há uma correlação positiva entre $\sum_i w_i$ e $\sigma(t_m)$. É perceptível como a condição $\sum_i w_i = 1$ resulta em previsões com pouca variabilidade, uma vez que o desvio padrão está, em média, muito próximo de 0.

Outro aspecto relevante ao determinar uma condição de normalização é avaliar o desempenho do estimador ao fazer previsões na base de teste. Na Figura 7, são apresentadas a métrica CI, conforme descrita na Equação 5.2, e o IBS, definido na Equação 5.6. Percebe-se não haver diferenças expressivas na métrica CI, sugerindo que a condição de normalização não interfere substancialmente na ordem correta dos tempos de falha estimados. No entanto, em relação ao IBS, a condição $\sum_i w_i = 100$ e EX parece ser a mais vantajosa, pois valores menores desta métrica indicam um melhor desempenho preditivo.

Figura 7: Métricas de performance do estimador (CI e IBS) para diferentes condições de normalizações $\sum_i w_i$ e funções de similaridade s_w



Fonte: Elaborada pela autora.

Neste ponto, podemos concluir que, entre as condições de normalização relatadas, $\sum_i w_i = 100$ parece ser a mais vantajosa. Portanto, a partir de agora, optamos por utilizar $\sum_i w_i = 100$ para as demais simulações.

6.1.2 Tempo Estimado de Falha

Na seção anterior, optamos por fixar o tempo estimado de falha como o t_m. Entretanto, esta não é a única alternativa, outra alternativa comumente descrita na literatura é utilizar o tempo mediano $t_{0.5}$ da curva de sobrevivência (Colosimo; Giolo, 2006). Esta escolha não interfere na estimativa de $\mathbf{w}$, tampouco no cálculo do IBS. Por este motivo, reportamos na Tabela 7 somente o CI e o desvio padrão dos tempos de falha estimados $\sigma(\hat{t})$.

Os resultados indicam que não há diferença considerável ao optar por qualquer uma das formas para estimar t. No entanto, encontramos uma dificuldade computacional ao calcular $t_{0.5}$, pois nem todas as curvas de sobrevivência estimadas decaem para a probabilidade de 0.5. Nessas situações, fomos obrigados a considerar o tempo mediano como o último tempo observado. Diante disso, para evitar soluções de contorno, decidimos fixar a estimativa de t como sendo o t_m.

Tabela 7: Métrica de avaliação CI e desvio padrão dos tempos de falha estimados $\sigma(\hat{t})$ utilizando o tempo médio t_m e o tempo mediano $t_{0.5}$ para diferentes funções de similaridade s_w

$\hat{t}$	s_w	CI	$\sigma(\hat{t})$
t_m	EX	0.772 (0.738, 0.805)	7.60 (6.69, 8.47)
	FR	0.773 (0.741, 0.806)	3.43 (3.06, 3.73)
$t_{0.5}$	EX	0.763 (0.727, 0.797)	8.13 (7.01, 9.33)
	FR	0.761 (0.725, 0.795)	3.49 (3.15, 3.82)

Fonte: Elaborada pela autora.

6.1.3 Distância Euclidiana Ponderada

Na base de crédito com a qual estamos trabalhando, lidamos exclusivamente com covariáveis numéricas, o que nos permite explorar diversas opções para calcular a distância entre essas covariáveis. Até o momento, implementamos apenas a Distância Euclidiana Ponderada, conforme descrita na Equação 3.2. No entanto, vamos agora considerar a inclusão de um novo parâmetro na DEP para aumentar sua flexibilidade e avaliar se isso pode levar a melhorias na capacidade do estimador de capturar as relações entre as variáveis.

Dessa forma, redefinimos a Distância Euclidiana Ponderada como:

$$d_w(\mathbf{x}_i, \mathbf{x}_j) = \left(\sum_{l=1}^{m} w_l (x_i^l - x_j^l)^2 \right)^q . \tag{6.1}$$

Observe que, dado que $\sum_{l=1}^{m} w_l (x_i^l - x_j^l)^2 > 1$, à medida que q aumenta, a distância definida em 6.1 também aumenta. Para $q = 1$, recuperamos a distância padrão.

Tabela 8: Pesos estimados w_1, w_2 e desvio padrão dos tempos de falha estimados $\sigma(t_m)$ para diferentes valores de q e funções de similaridade s_w

q	s_w	w_1	w_2	$\sigma(t_m)$
0.25	EX	52.27	47.72	1.64 (1.49, 1.77)
	FR	49.91	50.09	0.64 (0.60, 0.68)
0.5	EX	64.56	35.44	4.10 (4.27, 5.54)
	FR	56.24	43.75	1.44 (1.31, 1.55)
1.0	EX	75.65	24.35	7.60 (6.69, 8.47)
	FR	65.09	34.91	3.43 (3.06, 3.73)
2.0	EX	83.99	16.01	8.24 (7.32, 9.11)
	FR	74.22	25.78	6.60 (5.77, 7.35)
4.0	FR	82.22	17.78	7.97 (7.02, 8.83)

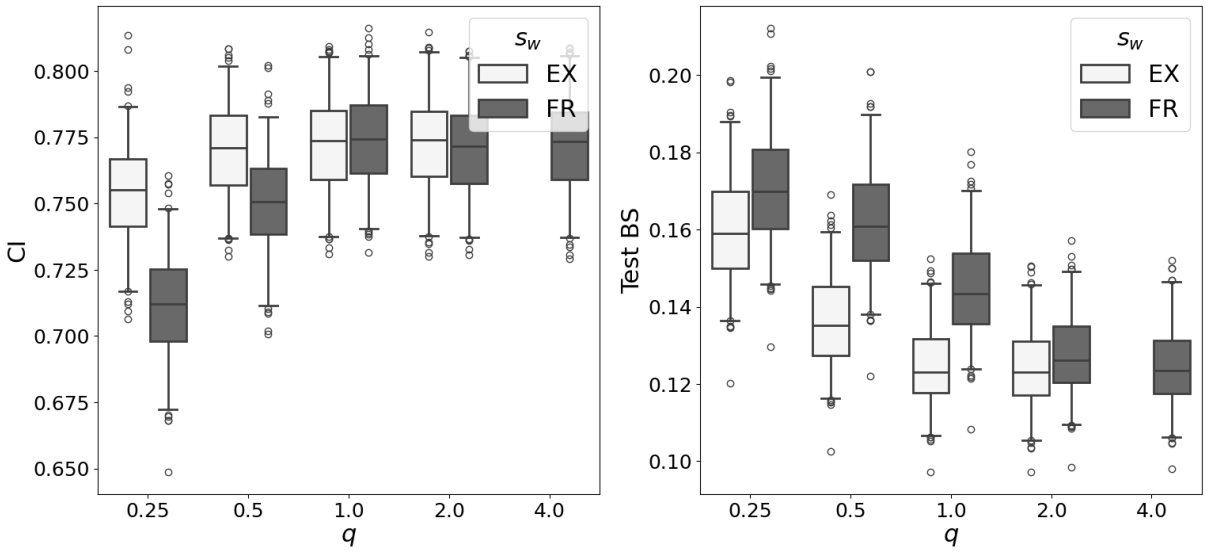
Fonte: Elaborada pela autora.

Inicialmente, é possível notar na Tabela 8 a ausência de resultados para $q = 4$ e EX. Isso decorre do fato de que a função de sobrevivência, calculada através da função de

similaridade exponencial, diverge rapidamente para grandes distâncias. Em outras palavras, à medida que a distância aumenta consideravelmente devido ao parâmetro q, a similaridade se aproxima mais rapidamente de 0 na função exponencial (EX) do que na fracionária (FR). Ao analisarmos a fórmula do SBKM na Equação 4.1, em algumas situações nos deparamos com uma divisão por um valor aproximadamente nulo que não conseguimos contornar.

Apesar dessa limitação, conseguimos obter resultados significativos. Por exemplo, é evidente uma correlação positiva entre o parâmetro q e a variabilidade dos tempos de falha estimados, indicada pelo desvio padrão σ. Além disso, reforçamos a observação de que a função EX produz previsões com maior variabilidade em comparação com a FR, e que a ordem de relevância entre as covariáveis se mantém em todos os cenários avaliados.

Figura 8: Métricas de performance do estimador (CI e IBS) para diferentes valores de q e funções de similaridade s_w



Fonte: Elaborada pela autora.

Ao analisar o desempenho do estimador na Figura 8, não identificamos diferenças relevantes na seleção do parâmetro q em relação ao CI. No entanto, observamos que o IBS parece ser um pouco sensível ao parâmetro no caso da função de similaridade FR.

6.1.4 Distância de Minkowski Ponderada

Por outro lado, a Distância de Minkowski Ponderada (DMP) atua como uma generalização para outras métricas de distância. Porém, é importante notar que, no estimador SBKM, estamos empregando a Distância Euclidiana quadrada, o que significa que a DEP não será um caso particular da DMP. Sua fórmula pode ser expressa da seguinte maneira:

$$d_w(\mathbf{x}_i, \mathbf{x}_j) = \left(\sum_{l=1}^{m} w_l |x_i^l - x_j^l|^p \right)^{1/p} \tag{6.2}$$

A DMP possibilita o cálculo da distância levando em conta a diferença em cada coordenada e a influência do parâmetro p na interpretação da proximidade dos pontos. Quando $p = 1$, resultamos na Distância de Manhattan, enquanto para $p = 2$, obtemos a Distância Euclidiana, correspondente a $q = 1/2$ na Equação 6.1.

Mais uma vez, enfrentamos dificuldades ao estimar a curva de sobrevivência utilizando a função de similaridade EX para casos onde $p < 1$, devido ao termo $1/p$. Ao analisar o desvio padrão dos tempos estimados de falha na Tabela 9, notamos uma correlação negativa com o valor de p.

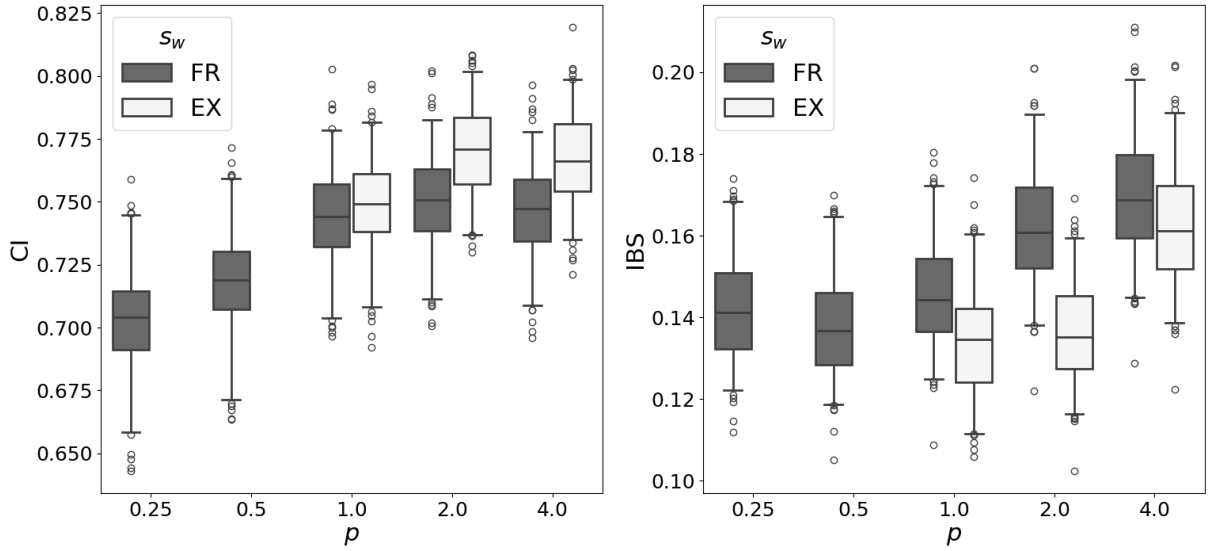
Tabela 9: Pesos estimados w_1, w_2 e desvio padrão dos tempos de falha estimados $\sigma(t_m)$ para diferentes valores de p e funções de similaridade s_w

p	s_w	w_1	w_2	$\sigma(t_m)$
0.25	FR	88.23	11.77	6.83 (6.15, 7.62)
0.5	FR	85.01	14.99	5.79 (5.24, 6.46)
1.0	EX	89.49	10.51	9.79 (8.57, 10.78)
	FR	66.74	33.26	3.28 (3.06, 3.51)
2.0	EX	64.56	35.44	4.91 (4.27, 5.55)
	FR	56.25	43.75	1.44 (1.31, 1.55)
4.0	EX	54.9	45.1	1.48 (1.29, 1.64)
	FR	53.58	46.42	0.72 (0.65, 0.78)

Fonte: Elaborada pela autora.

Com base na Figura 9, podemos deduzir que aumentar o valor de p parece ser benéfico para a qualidade da ordenação das previsões, já que a métrica CI aumenta utilizando ambas as funções de similaridade. No entanto, o IBS fornece um limite superior para p, pois os valores em $p = 4$ para esta métrica sugerem uma qualidade inferior nas previsões. Apesar disso, não existem diferenças notáveis nos resultados para afirmar com certeza.

Figura 9: Métricas de performance do estimador (CI e IBS) para diferentes valores de p e funções de similaridade s_w



Fonte: Elaborada pela autora.

Para facilitar a compreensão do impacto das diferentes distâncias nas métricas de avaliação, a Tabela 10 é apresentada. Observa-se que parece não haver diferença considerável em nenhuma das métricas relatadas. No entanto, optamos por manter a distância euclidiana ponderada (DEP), com $q = 1$, como a métrica de distância preferida para as próximas simulações, devido à sua simplicidade e ao seu amplo uso na literatura.

Tabela 10: Métricas de performance do estimador (CI e IBS) para diferentes distâncias definidas pelos valores de p e q, e funções de similaridade s_w

d_w		s_w	**CI**	**IBS**
DMP	$p = 0.25$	FR	0.702 (0.658, 0.745)	0.142 (0.122, 0.168)
	$p = 0.5$	FR	0.718 (0.671, 0.759)	0.138 (0.119, 0.165)
	$p = 1.0$	EX	0.748 (0.708, 0.781)	0.134 (0.112, 0.160)
		FR	0.744 (0.704, 0.778)	0.146 (0.125, 0.172)
	$p = 2.0$	EX	0.770 (0.737, 0.802)	0.136 (0.116, 0.159)
		FR	0.750 (0.711, 0.783)	0.162 (0.138, 0.190)
	$p = 4.0$	EX	0.767 (0.735, 0.798)	0.162 (0.139, 0.190)
		FR	0.746 (0.709, 0.778)	0.170 (0.145, 0.198)
DEP	$q = 0.25$	EX	0.754 (0.717, 0.787)	0.160 (0.136, 0.188)
		FR	0.712 (0.672, 0.748)	0.171 (0.146, 0.199)
	$q = 0.5$	EX	0.770 (0.737, 0.802)	0.136 (0.116, 0.159)
		FR	0.750 (0.711, 0.783)	0.162 (0.138, 0.190)
	$q = 1.0$	EX	0.772 (0.738, 0.805)	0.125 (0.107, 0.146)
		FR	0.773 (0.741, 0.806)	0.145 (0.124, 0.170)
	$q = 2.0$	EX	0.773 (0.738, 0.807)	0.124 (0.105, 0.146)
		FR	0.771 (0.737, 0.805)	0.128 (0.110, 0.149)
	$q = 4.0$	FR	0.772 (0.737, 0.806)	0.125 (0.106, 0.146)

Fonte: Elaborada pela autora.

6.1.5 Invariabilidade dos Pesos Estimados

Nesta seção, nosso objetivo é avaliar a invariabilidade das estimativas obtidas por meio do SBKM. Em primeiro lugar, destacamos a importância de obter consistentemente o mesmo resultado, independentemente do chute inicial, ao estimar parâmetros com um método de otimização. A repetibilidade do resultado proporciona confiança de que o método de otimização utilizado convergiu para a solução correta, independentemente das condições iniciais fornecidas. Isso é especialmente importante para as nossas análises, pois a precisão e a confiabilidade dos resultados são essenciais para a validade das conclusões obtidas. Portanto, realizamos um teste variando o chute inicial, gerado aleatoriamente dentro do intervalo de 0 a 100 utilizando um módulo integrado do Python para a criação de números aleatórios. Os resultados obtidos estão apresentados na Tabela 11. Observamos que a variação dos pesos estimados é mínima, e essa pequena diferença decimal não influencia nas métricas de desempenho avaliadas.

Tabela 11: Pesos estimados w_1, w_2 e métricas de performance do estimador (CI e IBS) para diferentes chutes iniciais

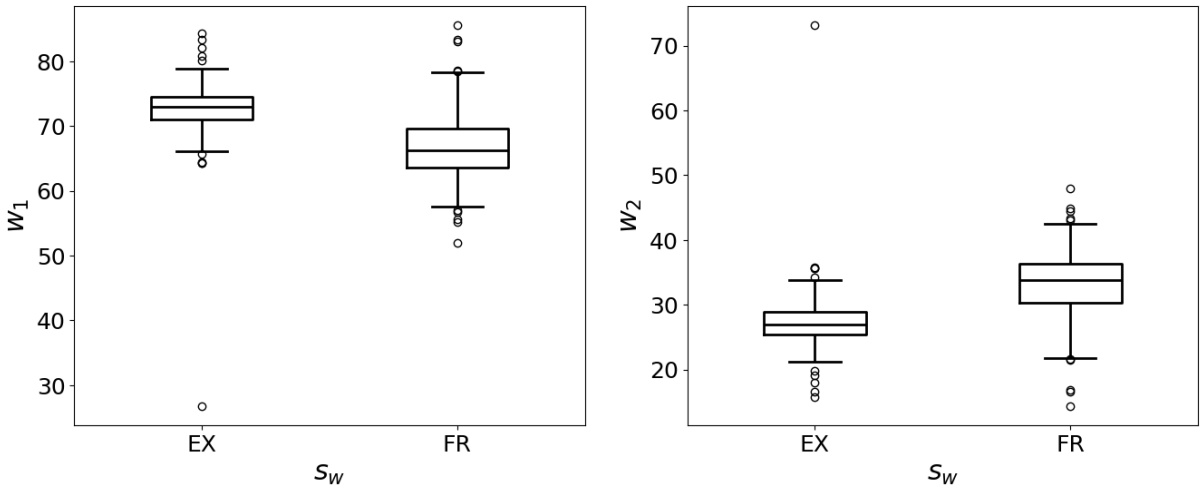
s_w	w_1	w_2	**CI**	**IBS**
EX	75.65 (75.64, 75.65)	24.35 (24.35, 24.36)	0.774 (0.774, 0.774)	0.125 (0.125, 0.125)
FR	65.09 (65.08, 65.09)	34.91 (34.91, 34.92)	0.775 (0.775, 0.775)	0.136 (0.136, 0.136)

Fonte: Elaborada pela autora.

Além disso, decidimos realizar uma repetição da técnica de amostragem *Bootstrap* com reposição na base de treino, visando uma avaliação mais genérica da consistência e invariabilidade dos pesos estimados. A partir dessa abordagem, estimamos os pesos para cada uma das amostras, permitindo a definição de intervalos de confiança não apenas para os pesos, mas também para as métricas de desempenho em treino e validação. Essa análise visa compreender a variação dos pesos estimados e das métricas de qualidade de previsão ao utilizar diferentes amostras de treinamento.

Ao analisar a Figura 10, nota-se que os pesos estimados utilizando a função de similaridade EX apresentam intervalos de confiança menores, indicando que esta função produz previsões mais consistentes, independentemente da base em que os pesos foram estimados. Além disso, não há diferença relevante entre os pesos obtidos por meio da função EX e da FR.

Figura 10: Pesos estimados w_1, w_2 para diferentes formas da função de similaridade s_w

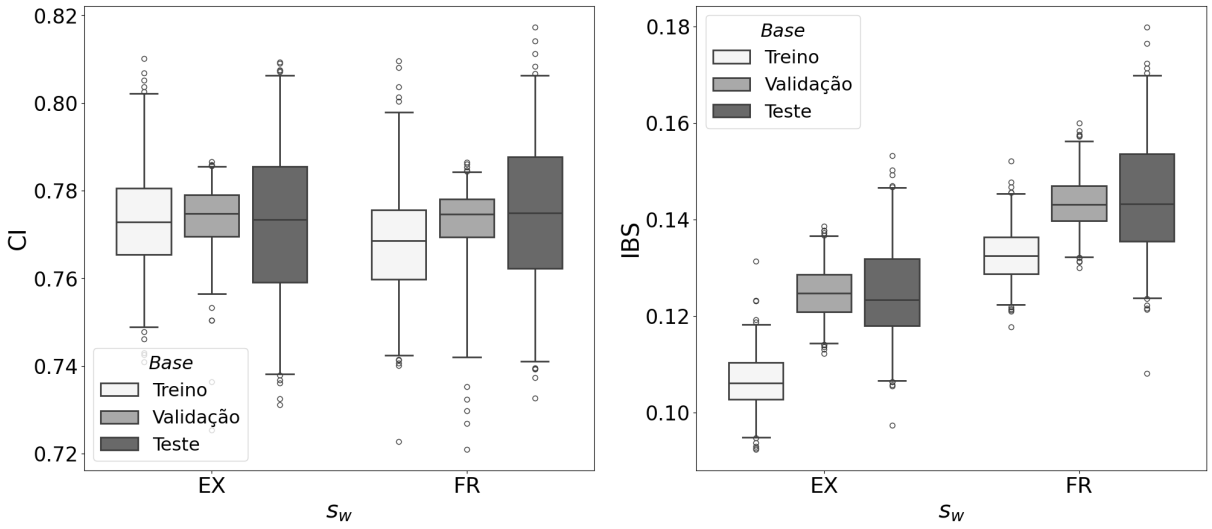


Fonte: Elaborada pela autora.

A partir da Figura 11, podemos avaliar o desempenho do estimador em diferentes bases de dados, separadas inicialmente antes da estimativa dos pesos. Observa-se que o desempenho do estimador em relação à métrica CI é consistente nas diferentes amostras, o que indica a eficácia do estimador em distinguir os indivíduos com maior ou menor risco. Por outro lado,

nota-se que o IBS apresenta ligeira melhoria na base de treinamento em comparação com a base de validação e teste. Adicionalmente, os resultados do IBS permitem identificar e sugerir que o desempenho da função EX é ligeiramente superior ao da função FR. Por esse motivo, preferimos fixar a função EX para os testes posteriores.

Figura 11: Métricas de performance do estimador (CI e IBS) em treino, validação e teste para diferentes funções de similaridade s_w



Fonte: Elaborada pela autora.

6.1.6 Modelos de Referência

Esta seção enfoca a avaliação de simulações realizadas em diversos modelos de referência descritos no Capítulo 5, comparativamente ao estimador proposto SBKM. A análise comparativa é crucial para avaliar a competitividade e eficácia de uma nova abordagem. Ao confrontar o desempenho de um novo modelo com uma variedade de modelos de referência bem estabelecidos, podemos obter entendimentos sobre sua capacidade de generalização, invariabilidade e desempenho em diferentes métricas de avaliação. Essa abordagem nos permite situar o desempenho do modelo proposto em relação ao estado-da-arte atual, fornecendo uma base sólida para avaliar sua utilidade e relevância em aplicações práticas. Através dessas comparações, é possível identificar as vantagens e limitações do nosso estimador, bem como áreas potenciais para melhorias futuras.

Todos os modelos comparativos foram avaliados nas mesmas condições do estimador SBKM. Isso implica que as mesmas covariáveis e o mesmo conjunto de dados de treino, validação e teste foram utilizados para cada modelo. Além disso, para cada modelo de referência, conduzimos uma otimização dos hiperparâmetros por meio de uma técnica denominada *Grid*

Search. Essa técnica envolve uma pesquisa exaustiva por todas as combinações possíveis dos valores de parâmetros especificados para um estimador. Mais detalhes sobre o espaço de busca e a otimização dos hiperparâmetros estão disponíveis no Apêndice A desta dissertação.

Dito isso, uma vez que o estimador SBKM funciona como uma adaptação e extensão do estimador de Kaplan-Meier, também fornecemos as métricas de desempenho deste último como referência de base. Isso se deve ao entendimento de que o SBKM deve, no mínimo, superar o desempenho do KM. Portanto, o estimador de KM estima a mesma curva de sobrevivência para todos os indivíduos da base de teste e, consequentemente, o mesmo tempo estimado de falha.

Tabela 12: Métricas de performance dos modelos (CI e IBS) em treino, validação e teste

	Treino		**Validação**		**Teste**	
Modelo	**CI**	**IBS**	**CI**	**IBS**	**CI**	**IBS**
KM	-	-	-	-	0.500 (0.500, 0.500)	0.170 (0.146, 0.198)
SBKM	0.770	0.111	0.777	0.121	0.772 (0.738, 0.805)	0.125 (0.107, 0.146)
COX	0.770	0.123	0.788	0.116	0.775 (0.738, 0.808)	0.118 (0.102, 0.137)
EN-COX	0.769	0.123	0.789	0.117	0.774 (0.739, 0.808)	0.118 (0.102, 0.138)
WEIBULL	0.770	0.121	0.789	0.113	0.776 (0.735, 0.806)	0.115 (0.098, 0.139)
ST	0.764	0.124	0.789	0.128	0.722 (0.688, 0.758)	0.130 (0.110, 0.155)
RST	0.789	0.118	0.793	0.122	0.762 (0.724, 0.798)	0.128 (0.110, 0.151)
GB-COX	0.873	0.080	0.807	0.121	0.726 (0.681, 0.767)	0.137 (0.113, 0.166)

Fonte: Elaborada pela autora.

Ao analisar a Tabela 12, é evidente que o SBKM se mostra competitivo em relação aos demais modelos, demonstrando até mesmo uma ligeira superioridade em relação aos modelos ST e GB-COX em ambas as métricas de avaliação na base de teste. Contudo, é importante notar que tanto no modelo ST quanto no GB-COX, foi observado um problema de sobreajuste. Este fenômeno é comum em modelos de árvore que são muito flexíveis e se ajustam excessivamente aos dados de treinamento (Song; Lu, 2015), podendo resultar em perda de generalização, especialmente em bases de dados pequenas, como a base de risco de crédito utilizada neste estudo. Existem estratégias que podem ser implementadas para mitigar esses problemas, mas tais técnicas estão além do escopo deste trabalho (Vezhnevets; Barinova, 9007). Outro aspecto relevante é que, no contexto do SBKM, há uma facilidade na interpretação dos parâmetros, o que nem sempre ocorre ao utilizar modelos complexos de aprendizado de máquina.

Em contrapartida, os modelos estatísticos de análise de sobrevivência, como COX, EN-COX e WEIBULL, apresentaram um desempenho levemente superior na base de teste em relação ao IBS, quando comparados ao estimador SBKM e todos os outros modelos de

aprendizado de máquina. Este resultado ressalta a validade e a qualidade dos modelos estatísticos como referência para a resolução de problemas envolvendo análise de sobrevivência. No entanto, é importante notar que essa diferença não parece ser significativa, o que nos leva a concluir que o SBKM também é uma abordagem viável para estimar curvas de sobrevivência condicionais. Além disso, uma vantagem do SBKM é que não é necessário fazer suposições sobre distribuições de probabilidade ou riscos proporcionais.

6.1.7 Amostragem

Um desafio significativo ao empregar modelos de sobrevivência baseados em similaridade na prática é a dependência desses modelos do conhecimento prévio da similaridade entre uma amostra de teste e cada amostra de treinamento. A computação necessária para implementar esses modelos pode se tornar inviável à medida que o tamanho do conjunto de dados aumenta. Essa limitação é particularmente característica em modelos baseados em instâncias, onde são armazenadas todas as instâncias de treinamento para a classificação de uma nova instância (Wilson; Martinez, 2000). No caso do estimador SBKM, essa dependência é evidenciada pela matriz de similaridade definida na Equação 4.9.

Embora o tamanho da amostra não seja um problema para a base de crédito, gostaríamos de avaliar o desempenho do estimador quando reduzimos o tamanho da base de treinamento. Nosso objetivo é compreender como o desempenho do estimador se mantém à medida que diminuímos os dados utilizados para estimar os pesos. Portanto, mantemos a base de validação e teste inalteradas para garantir uma análise comparativa consistente com os demais cenários. Para esta análise, optamos por fixar a função de similaridade na sua forma exponencial (EX).

Na Tabela 13, é possível notar que, para o menor valor de n, os desempenhos em treino são os melhores. Isso ocorre porque é mais fácil para os pesos se ajustarem adequadamente a um conjunto pequeno de dados. Por outro lado, os desempenhos em teste são os piores, pois o estimador não tem a capacidade de generalização necessária para fazer boas inferências em dados desconhecidos. Além disso, é notável que, a partir de $n = 100$, as métricas alcançadas são tão boas quanto quando se utiliza a base de treinamento completa para estimar os pesos. Assim, concluímos que o estimador se adapta bem a uma redução no tamanho da amostra para realizar previsões. Por outra perspectiva, observamos que aumentar o tamanho da amostra além de $n = 100$ não parece resultar em melhorias significativas no desempenho.

Tabela 13: Pesos estimados w_1, w_2 e métricas de performance do estimador (CI e IBS) em treino, validação e teste para amostras de diferentes tamanhos n

			Treino		**Validação**		**Teste**	
n	w_1	w_2	**CI**	**IBS**	**CI**	**IBS**	**CI**	**IBS**
25	61.55	38.45	0.888	0.034	0.646	0.121	0.605 (0.551, 0.655)	0.247 (0.212,0.278)
50	63.39	36.61	0.778	0.115	0.644	0.260	0.643 (0.589, 0.690)	0.173 (0.155, 0.192)
100	66.09	33.91	0.811	0.108	0.752	0.174	0.751 (0.712, 0.784)	0.127 (0.115, 0.143)
250	70.74	29.26	0.755	0.108	0.776	0.137	0.762 (0.727, 0.794)	0.121 (0.103, 0.143)
560	75.65	24.35	0.770	0.111	0.777	0.121	0.772 (0.738, 0.805)	0.125 (0.107, 0.146)

Fonte: Elaborada pela autora.

Com base nisso, realizamos uma simulação comparativa semelhante à conduzida na seção anterior, conforme demonstrado na Tabela 14, para investigar se outros modelos exibem a mesma característica que o estimador SBKM nessas circunstâncias. Observamos que o desempenho de todos os modelos para uma amostra de tamanho $n = 250$ foi semelhante à performance obtida ao utilizar o conjunto de dados original. No entanto, o modelo ST apresentou uma piora média de aproximadamente 8.86% na métrica de CI em teste.

Tabela 14: Métricas de performance dos modelos (CI e IBS) em treino, validação e teste para uma amostra de $n = 250$

	Treino		**Validação**		**Teste**	
Modelo	**CI**	**IBS**	**CI**	**IBS**	**CI**	**IBS**
KM	-	-	-	-	0.500 (0.500, 0.500)	0.179 (0.157, 0.206)
SBKM	0.755	0.108	0.776	0.128	0.762 (0.737, 0.794)	0.121 (0.103, 0.143)
COX	0.758	0.115	0.788	0.120	0.773 (0.738, 0.807)	0.118 (0.104, 0.137)
EN-COX	0.757	0.114	0.788	0.121	0.773 (0.737, 0.807)	0.119 (0.104, 0.136)
WEIBULL	0.758	0.112	0.789	0.118	0.773 (0.738, 0.805)	0.116 (0.099, 0.136)
ST	0.681	0.676	0.116	0.131	0.658 (0.612, 0.694)	0.136 (0.115, 0.160)
RST	0.804	0.110	0.763	0.126	0.741 (0.704, 0.776)	0.127 (0.113, 0.144)
GB-COX	0.871	0.079	0.744	0.141	0.709 (0.668, 0.748)	0.140 (0.118, 0.166)

Fonte: Elaborada pela autora.

Para examinar a consistência dos pesos estimados e a performance do estimador SBKM em diferentes amostras aleatórias de tamanho $n = 250$, repetimos a técnica de *Bootstrap* sem reposição. Construímos intervalos de confiança com 95% de confiabilidade para os pesos estimados, conforme mostrado na Tabela 15, e para as métricas de treino, validação e teste, como ilustrado na Figura 12.

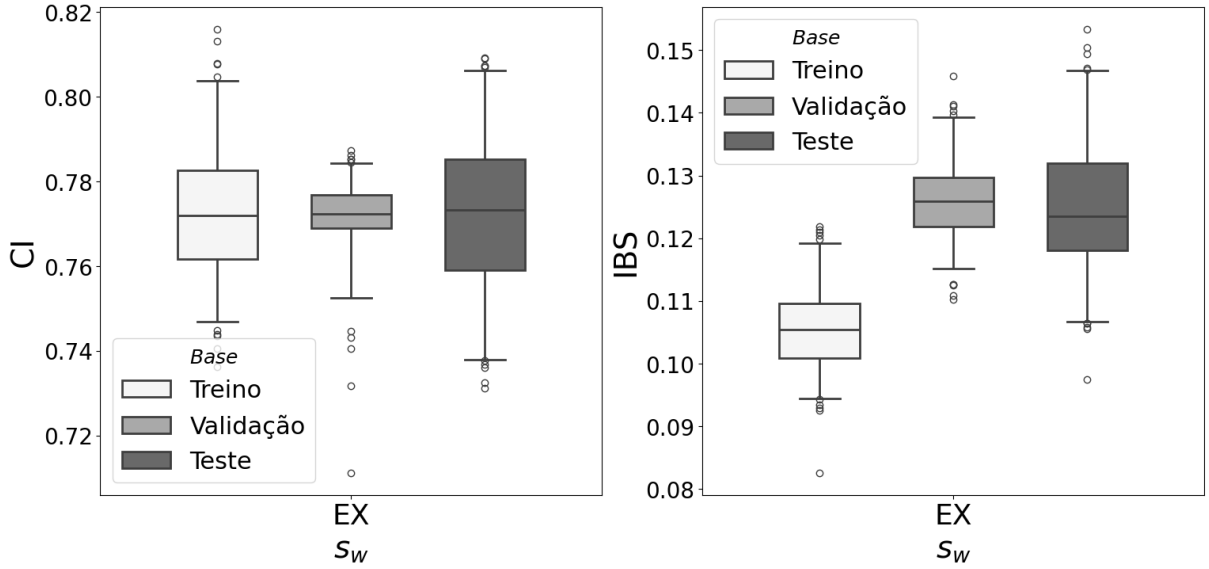
Podemos inferir que há uma consistência semelhante tanto nos pesos quanto nas métricas de avaliação quando os pesos são estimados com uma amostra reduzida de $n = 250$, em comparação com a amostra de $n = 560$.

Tabela 15: Pesos estimados w_1, w_2 com seus respectivos intervalos de confiança para amostras de $n = 250$ e $n = 560$

n	w_1	w_2
250	71.93 (67.24, 76.40)	28.07 (23.60, 32.76)
560	72.68 (66.16, 78.88)	27.32 (21.12, 33.84)

Fonte: Elaborada pela autora.

Figura 12: Métricas de performance do estimador (CI e IBS) em treino, validação e teste para amostras de tamanho $n = 250$



Fonte: Elaborada pela autora.

6.1.8 Taxa de Censura

A taxa de censura em uma amostra desempenha um papel crucial na análise de sobrevivência, pois pode ter um impacto direto na precisão das estimativas e interpretações dos resultados. Quanto maior a taxa de censura, menos informação está disponível sobre os tempos de falha reais, o que pode levar a estimativas menos precisas dos parâmetros. Para investigar esse efeito, realizamos uma simulação fixando o tamanho da base de treinamento em $n = 250$ e selecionando aleatoriamente, sem repetição, falhas e censuras para obter diferentes taxas de falha. Em seguida, avaliamos os pesos estimados, o desvio padrão dos tempos de falha estimados e o desempenho nas métricas de avaliação na base de teste.

Observamos na Tabela 16 que a menor taxa de censura apresenta, em média, o menor desvio padrão dos tempos de falha estimados. Por outro lado, o maior desvio padrão é observado quando metade da amostra é composta por censuras.

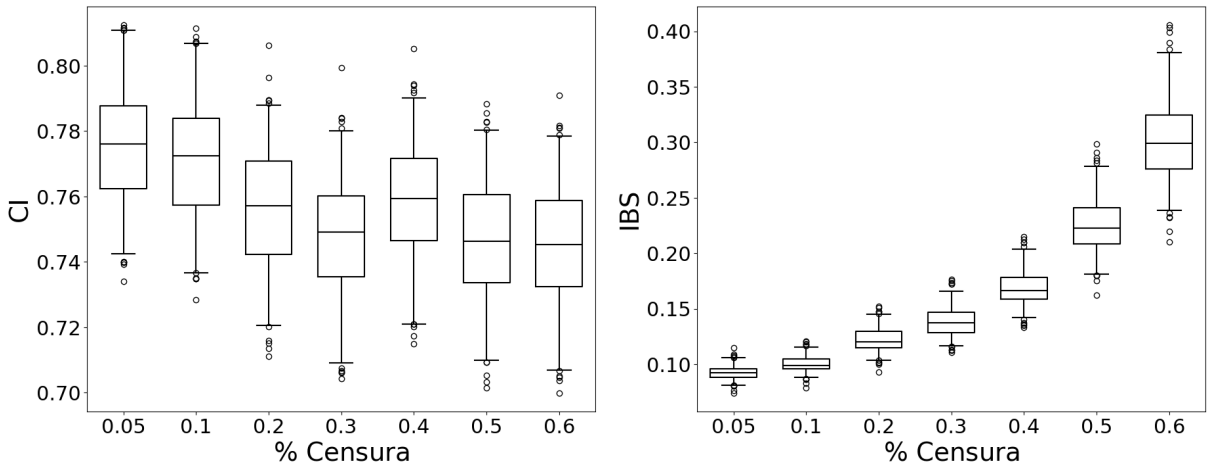
Tabela 16: Pesos estimados w_1, w_2 e desvio padrão dos tempos de falha estimados $\sigma(t_m)$ para amostras com diferentes taxas de censura

% Censura	w_1	w_2	$\sigma(t_m)$
0.05	78.90	21.10	6.18 (5.12, 7.15)
0.1	76.43	23.57	6.44 (5.37, 7.50)
0.2	70.83	29.17	7.21 (6.08, 8.31)
0.3	72.91	27.09	8.43 (7.25, 9.48)
0.4	70.93	29.07	8.60 (7.57, 9.63)
0.5	70.47	29.53	8.96 (8.00, 9.85)
0.6	73.46	26.54	8.60 (7.60, 9.38)

Fonte: Elaborada pela autora.

Com base nos resultados apresentados na Figura 13, podemos concluir que não foram observadas diferenças expressivas na métrica CI, indicando que a qualidade na discriminação dos indivíduos com maior ou menor risco permanece a mesma. No entanto, observamos que a medida de IBS é diretamente influenciada pela taxa de censura da base na qual os pesos foram estimados, resultando em uma piora significativa nesta métrica.

Figura 13: Métricas de performance do estimador (CI e IBS) para amostras com diferentes taxas de censura



Fonte: Elaborada pela autora.

6.2 Base de dados SUPPORT

Realizamos uma exploração em uma base de dados alternativa relacionada a um estudo clínico sobre o prognóstico de pacientes hospitalizados. A base foi dividida em conjuntos de treino, validação e teste, contendo 4968, 1243 e 2662 amostras, respectivamente. As covariáveis selecionadas foram “sod”, associada ao peso w_1, e “ca”, associada ao peso w_2. “sod” é uma variável numérica, enquanto “ca” é uma variável categórica nominal. Portanto, nestas condições,

utilizamos tanto a DEP como a DBP para o cálculo de distância.

Devido ao tamanho da base de dados SUPPORT, encontramos desafios para realizar simulações com o mesmo nível de detalhe em comparação com a base de dados CREDIT. Portanto, diferentes testes realizados anteriormente não serão repetidos para esta nova base. Embora estejamos cientes das limitações computacionais do estimador SBKM, possíveis soluções e estratégias para contorná-las não serão abordadas neste trabalho, mas foram trabalhadas por Wang *et al.* (2019) no contexto de estimadores baseados em kernel.

6.2.1 Tempo Estimado de Falha

De forma semelhante à análise conduzida na base de crédito, investigamos o impacto de calcular o tempo estimado de falha de cada indivíduo a partir do tempo médio t_m e do tempo mediano $t_{0.5}$ da curva de sobrevivência na Tabela 17. Notamos que não há diferença substancial na escolha de qualquer uma das formas em relação à métrica de avaliação CI. No entanto, observamos que a utilização de $t_{0.5}$ parece resultar em uma menor variabilidade nos tempos de falha. Além disso, mais uma vez, agora para a nova base de dados, constatamos a partir do desvio padrão $\sigma(\hat{t})$ que a função de similaridade na sua forma exponencial (EX) tende a produzir tempos de falha com maior variabilidade em comparação com a utilização na forma fracionária (FR).

Tabela 17: Métrica de avaliação CI e desvio padrão dos tempos de falha estimados $\sigma(\hat{t})$ utilizando o tempo médio t_m e o tempo mediano $t_{0.5}$ para diferentes funções de similaridade s_w

$\hat{t}$	s_w	CI	$\sigma(\hat{t})$
t_m	EX	0.571 (0.559, 0.585)	226.13 (222.18, 229.69)
	FR	0.571 (0.558, 0.585)	190.76 (187.34, 194.13)
$t_{0.5}$	EX	0.570 (0.556, 0.584)	184.18 (181.69, 186.51)
	FR	0.569 (0.555, 0.584)	160.49 (158.45, 162.55)

Fonte: Elaborada pela autora.

6.2.2 Modelos de Referência

Na Tabela 18, é possível avaliar o desempenho comparativo de cada um dos modelos em relação ao estimador SBKM em suas duas principais formas: EX e FR. Observa-se que, aparentemente, a base SUPPORT parece apresentar maior complexidade de modelagem em comparação com a base CREDIT, uma vez que não há diferença significativa na métrica de qualidade IBS entre a curva de sobrevivência estimada pelo Kaplan-Meier e pelos outros modelos,

incluindo o estimador SBKM. No entanto, devido à falta de capacidade discriminatória do KM para distinguir os usuários em baixo e alto risco, os demais modelos mostram-se mais eficazes na métrica de CI. Todos os modelos de referência analisados e o SBKM apresentaram desempenho semelhante em ambas as métricas observadas.

Neste contexto, é pertinente ressaltar que, devido à presença de uma covariável categórica, o SBKM se destaca, pois é o único modelo no qual não é necessário realizar a codificação dessa variável. Utilizamos apenas uma medida de distância apropriada para variáveis categóricas, permitindo avaliar a similaridade entre elas. Nos demais modelos, são criadas variáveis indicadoras, uma para cada categoria da covariável. Após essa codificação, as variáveis indicadoras são tratadas como se fossem completamente distintas, sem qualquer relação entre si. Assim, o estimador proposto considera um número menor de parâmetros, tornando-o mais parcimonioso.

Tabela 18: Métricas de performance dos modelos (CI e IBS) em treino, validação e teste

	Treino		**Validação**		**Teste**	
Modelo	**CI**	**IBS**	**CI**	**IBS**	**CI**	**IBS**
KM	-	-	-	-	0.500 (0.500, 0.500)	0.219 (0.208, 0.230)
SBKM (EX)	0.565	0.203	0.592	0.211	0.571 (0.559, 0.585)	0.206 (0.196, 0.216)
SBKM (FR)	0.563	0.204	0.589	0.212	0.571 (0.558, 0.585)	0.206 (0.197, 0.217)
COX	0.556	0.204	0.563	0.212	0.564 (0.553, 0.579)	0.207 (0.197, 0.217)
EN-COX	0.555	0.204	0.566	0.212	0.561 (0.551, 0.574)	0.207 (0.197, 0.217)
WEIBULL	0.556	0.205	0.563	0.214	0.564 (0.551, 0.577)	0.208 (0.199, 0.217)
ST	0.568	0.202	0.593	0.213	0.569 (0.558, 0.582)	0.207 (0.196, 0.216)
RST	0.568	0.202	0.596	0.211	0.569 (0.556, 0.582)	0.206 (0.196, 0.216)
GB-COX	0.571	0.203	0.600	0.212	0.572 (0.560, 0.585)	0.206 (0.196, 0.216)

Fonte: Elaborada pela autora.

6.2.3 Amostragem

A principal dificuldade encontrada ao lidar com a base SUPPORT foi sua dimensão. Como mencionado anteriormente, o SBKM enfrenta limitações computacionais em bases de dados extensas, pois sua fórmula, por definição, requer o cálculo da similaridade entre todos os indivíduos da amostra. Portanto, decidimos realizar análises em subamostras aleatórias para avaliar a qualidade das previsões feitas. Nesse contexto, para simplificar, optamos por continuar com a função de similaridade EX.

Na Tabela 19, é notável que os pesos estimados para amostras de diferentes tamanhos foram muito semelhantes, variando apenas nas casas decimais. Além disso, ambas as métricas

de desempenho apresentaram valores similares àqueles obtidos utilizando a maior base possível de treinamento, $n = 4968$. Esses resultados sugerem que não há diferença significativa no desempenho do estimador ao utilizar os diferentes tamanhos de amostra avaliados para estimar os pesos.

Tabela 19: Pesos estimados w_1, w_2 e métricas de performance do estimador (CI e IBS) em treino, validação e teste para amostras de diferentes tamanhos n

			Treino		**Validação**		**Teste**	
n	w_1	w_2	**CI**	**IBS**	**CI**	**IBS**	**CI**	**IBS**
100	93.91	6.09	0.632	0.202	0.551	0.219	0.555 (0.543, 0.569)	0.218 (0.210, 0.227)
250	93.78	6.22	0.597	0.190	0.559	0.205	0.565 (0.554, 0.579)	0.200 (0.191, 0.209)
500	93.86	6.14	0.562	0.202	0.557	0.215	0.568 (0.557, 0.581)	0.207 (0.198, 0.217)
1000	93.88	6.12	0.553	0.201	0.577	0.220	0.560 (0.546, 0.573)	0.211 (0.200, 0.221)
4968	93.62	6.38	0.565	0.202	0.592	0.211	0.571 (0.558, 0.585)	0.206 (0.196, 0.216)

Fonte: Elaborada pela autora.

Ao fixarmos o tamanho em $n = 250$, realizamos uma simulação comparativa entre modelos utilizando a mesma subamostra, conforme mostrado na Tabela 20. Observa-se que, na base de teste, em média, o SBKM se destaca positivamente em relação à métrica de CI, obtendo a maior média entre os modelos de referência analisados. No entanto, em relação ao IBS, a maioria dos modelos manteve uma qualidade similar àquela observada na base completa de treinamento original.

Tabela 20: Métricas de performance dos modelos (CI e IBS) em treino, validação e teste para uma amostra de $n = 250$

	Treino		**Validação**		**Teste**	
Modelo	**CI**	**IBS**	**CI**	**IBS**	**CI**	**IBS**
KM	-	-	-	-	0.500 (0.500, 0.500)	0.208 (0.198, 0.218)
SBKM	0.597	0.190	0.559	0.205	0.565 (0.552, 0.579)	0.200 (0.191, 0.209)
COX	0.586	0.199	0.561	0.204	0.562 (0.550, 0.576)	0.199 (0.191, 0.209)
EN-COX	0.586	0.199	0.562	0.203	0.562 (0.551, 0.577)	0.199 (0.191, 0.209)
WEIBULL	0.587	0.200	0.562	0.204	0.562 (0.548, 0.575)	0.200 (0.192, 0.207)
ST	0.632	0.182	0.534	0.223	0.545 (0.531, 0.558)	0.215 (0.206, 0.225)
RST	0.601	0.190	0.548	0.211	0.558 (0.546, 0.570)	0.205 (0.196, 0.215)
GB-COX	0.621	0.188	0.553	0.207	0.555 (0.542, 0.568)	0.203 (0.194, 0.214

Fonte: Elaborada pela autora.

Ademais, ao analisarmos a consistência dos pesos estimados pelo SBKM em diferentes amostras de tamanho $n = 250$, observamos na Tabela 21 que os pesos variam muito pouco em relação à média, e os pesos estimados para $n = 4956$, conforme mostrado na Tabela 19, estão

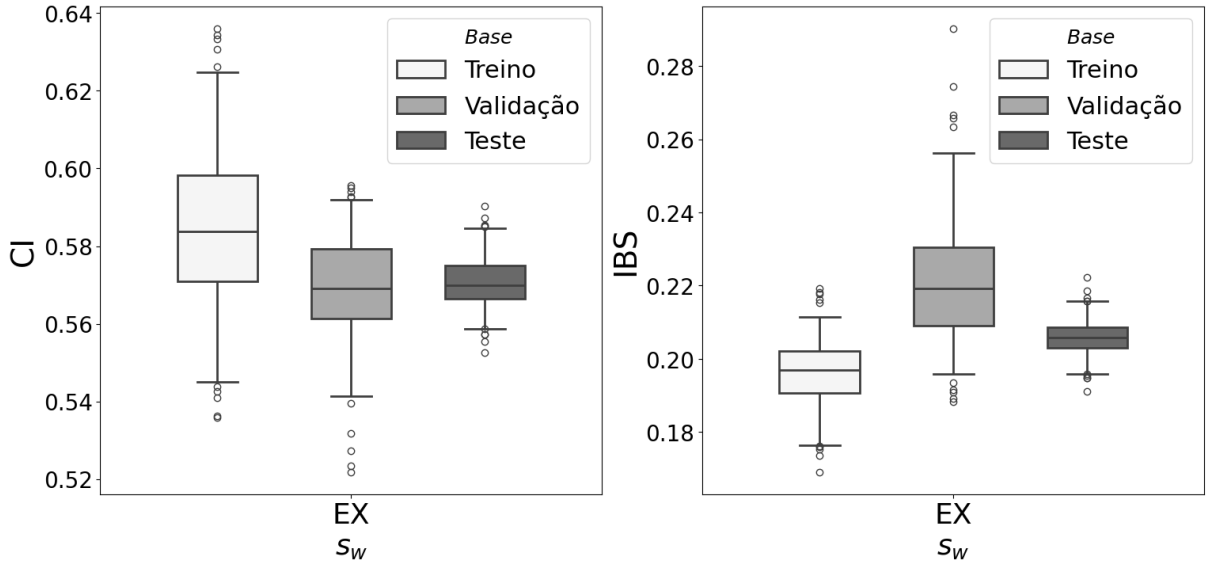
dentro do intervalo de confiança. No entanto, na Figura 14, percebemos como essas pequenas alterações nos pesos influenciam a variação da métrica de CI na base de treinamento, embora o intervalo de confiança seja bastante estreito na base de teste. Quanto à métrica de IBS, notamos que a maior variação de desempenho ocorre na base de validação.

Tabela 21: Pesos estimados w_1, w_2 com seus respectivos intervalos de confiança para amostras de $n = 250$

n	w_1	w_2
250	93.66 (93.39, 93.86)	6.33 (6.13, 6.60)

Fonte: Elaborada pela autora.

Figura 14: Métricas de performance do estimador (CI e IBS) em treino, validação e teste para amostras de tamanho $n = 250$



Fonte: Elaborada pela autora.

6.2.4 Taxa de Censura

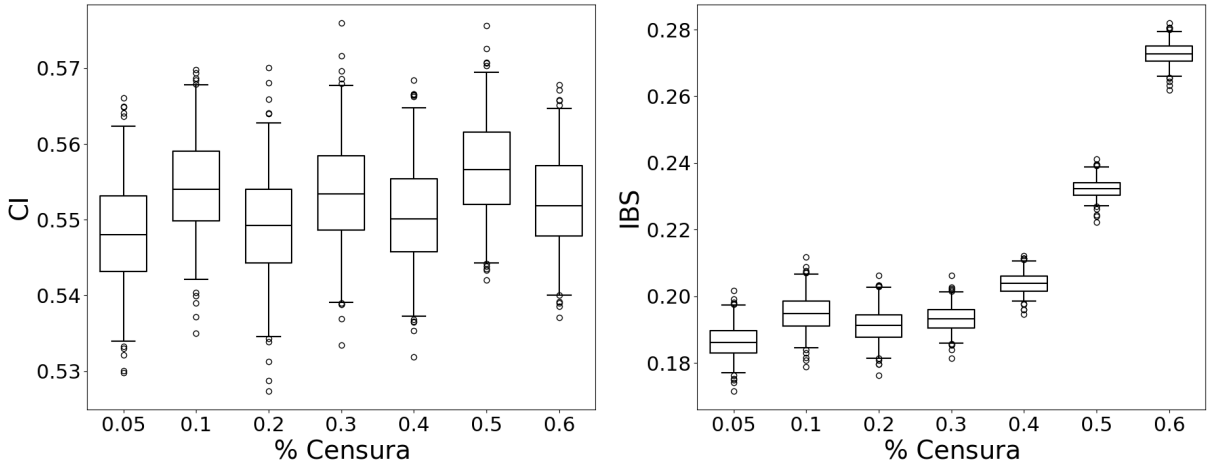
Por fim, para examinar a variação da taxa de censura na base de treinamento, mantivemos $n = 250$ e selecionamos aleatoriamente as falhas e censuras para alcançar a porcentagem desejada. Os resultados são mostrados na Tabela 22 e na Figura 15. Observa-se que quanto maior a taxa de censura, maior o desvio padrão dos tempos de falha estimados, sugerindo uma correlação positiva. As demais conclusões foram semelhantes às obtidas na base CREDIT avaliada na seção anterior: notamos que a métrica CI não parece ser afetada pela taxa de censura, enquanto o IBS apresenta uma piora relevante com o aumento de amostras censuradas na base.

Tabela 22: Pesos estimados w_1, w_2 e desvio padrão dos tempos de falha estimados $\sigma(t_m)$ para amostras com diferentes taxas de censura

% Censura	w_1	w_2	$\sigma(t_m)$
5	94.02	5.98	84.82 (81.78, 88.08)
10	93.90	6.10	109.21 (105.46, 112.09)
20	93.83	6.17	146.62 (142.14, 151.18)
30	93.85	6.15	208.33 (200.61, 216.65)
40	93.84	6.16	234.99 (228.40, 242.62)
50	93.82	6.18	265.25 (257.93, 274.36)
60	93.87	6.13	285.28 (275.26, 295.30)

Fonte: Elaborada pela autora.

Figura 15: Métricas de performance do estimador (CI e IBS) para amostras com diferentes taxas de censura



Fonte: Elaborada pela autora.

7 CONCLUSÕES

Neste trabalho, apresentamos uma adaptação do estimador de Kaplan-Meier (KM), denominada estimador de Kaplan-Meier baseado em similaridade (SBKM), no qual incorporamos uma função de similaridade em sua fórmula original. Essa nova abordagem possibilita a inclusão da influência das covariáveis na estimativa das curvas de sobrevivência, o que não é diretamente alcançável com o KM convencional. Assim, por meio do SBKM, podemos estimar curvas de sobrevivência condicionais para cada indivíduo. Conduzimos análises em duas bases de dados reais: CREDIT, que avalia mutuários e inadimplência, e SUPPORT, que avalia pacientes e óbito. Implementamos um método de seleção variáveis, no qual usamos o desempenho do estimador na base de validação como critério de decisão. Na base de dados CREDIT, foram consideradas duas covariáveis numéricas. Na base de dados SUPPORT, também foram consideradas duas covariáveis, sendo uma categórica nominal e a outra numérica.

Devido ao tamanho reduzido da base de dados CREDIT, pudemos explorar com facilidade diversas propriedades e hiperparâmetros do estimador. Inicialmente, conduzimos simulações avaliando as duas possibilidades da função de sobrevivência: EX e FR. Em primeiro lugar, avaliamos a condição de normalização dos pesos estimados, onde observamos que essa escolha parece ser determinante para a variabilidade dos tempos de falha estimados. Também fixamos $\sum_i w_i = 100$, pois observamos, com base nos resultados das métricas de avaliação CI e IBS, que obtivemos, em média, os melhores resultados.

Em seguida, investigamos a escolha entre o tempo médio e o tempo mediano da curva de sobrevivência para estimar o tempo de falha dos indivíduos, tanto no cálculo do CI quanto na variabilidade dos tempos estimados. Nesta análise, não encontramos diferenças significativas em nenhum dos dois cenários estudados. Por outro lado, os resultados sugerem, com base na maioria dos casos avaliados, que a função de similaridade EX gera estimativas com maior variabilidade em relação à FR. Essa conclusão se mantém também em todos os cenários avaliados posteriormente. Reiteramos que, de modo geral, buscamos estimativas com maior variabilidade, pois esperamos previsões mais sensíveis às variações nas covariáveis.

Além disso, exploramos diferentes medidas de distância: DEP e DMP. No entanto, em relação às métricas de avaliação, não observamos diferenças significativas em nenhuma das configurações avaliadas. Por esse motivo, e também por questões de simplicidade, decidimos fixar a métrica de distância utilizada como sendo a DEP com $q = 1$.

Para examinar a capacidade de invariabilidade dos parâmetros estimados, avaliamos

os pesos para diferentes conjuntos de treinamento utilizando a técnica de reamostragem *bootstrap*. Observamos que os pesos estimados usando a função EX parecem ser mais consistentes em comparação com a função FR. Além disso, em termos de IBS, a função EX produz, em média, previsões mais precisas do que a função FR. Com base nisso, optamos por fixar a função de similaridade usando sua versão exponencial (EX).

Conduzimos uma análise comparativa do estimador SBKM em relação a diferentes modelos estatísticos e de aprendizado de máquina, todos testados em cenários semelhantes. Isto é, mesma amostra e após uma breve otimização de seus hiperparâmetros. Os resultados obtidos pelo estimador proposto nas métricas avaliadas de CI e IBS foram competitivos em comparação com os demais modelos. Em particular, destacamos que o SBKM se diferencia por não fazer suposições sobre distribuições ou riscos proporcionais em comparação aos modelos estatísticos. Além disso, o SBKM possui parâmetros que poderiam ser mais facilmente interpretáveis e não observamos problemas de sobreajuste, ao contrário de alguns algoritmos de aprendizado de máquina utilizados.

Ademais, realizamos testes de desempenho do estimador em amostras de diferentes tamanhos para o conjunto de treinamento, mantendo as bases de validação e teste constantes. Concluímos que somente para as amostras de tamanho $n < 100$, houve uma piora significativa em ambas as métricas avaliadas. Para comparar o desempenho do SBKM com outros métodos de estimação neste mesmo cenário, mantivemos $n = 250$ e avaliamos as métricas de desempenho. Observamos que o único modelo que apresentou um desempenho significativamente inferior na métrica de CI foi a árvore de sobrevivência (ST). Ao final, constatamos que não houve diferença significativa nos pesos estimados utilizando $n = 560$, correspondente à base de treinamento definida no início do estudo, e $n = 250$.

Finalmente, avaliamos o desempenho do SBKM em amostras com diferentes taxas de censura. Para isso, mantivemos o tamanho da amostra fixo em $n = 250$ e selecionamos aleatoriamente, sem repetição, falhas e censuras com o objetivo de atingir a porcentagem predefinida. Exploramos uma faixa de taxa de censura de 5% a 60% e, dentro desse intervalo, não observamos diferenças significativas na métrica de CI. No entanto, notamos uma significativa piora na métrica de IBS à medida que o número de censuras na amostra aumenta, sugerindo uma correlação negativa.

No que diz respeito ao conjunto de dados SUPPORT, realizamos simulações semelhantes. Entretanto, devido ao tamanho da base de dados, nos deparamos com limitações

desafiadoras de tempo computacional, o que nos restringiu a conduzir apenas parte das análises. Inicialmente, examinamos a influência da escolha entre o tempo médio e o tempo mediano para estimar o tempo de falha na métrica CI, bem como na variabilidade dos tempos de falha. Assim como na base explorada anteriormente, não observamos diferenças significativas nessas análises.

Ao comparar o método proposto com outros modelos de referência, notamos que a base de dados SUPPORT parece ser mais complexa para todos os métodos, conforme evidenciado pelo CI próximo de 0.5, o que indica uma ordenação aleatória dos tempos de falha. Nessa análise, não encontramos diferenças significativas em nenhuma das métricas de desempenho avaliadas, reforçando a competitividade do SBKM neste contexto também. Salientamos ainda que, neste cenário em que há uma covariável categórica, o SBKM apresenta a vantagem de ser facilmente adaptável e mais parcimonioso em comparação com os outros modelos. Por fim, não observamos vantagens em utilizar a função de similaridade EX ou FR. Por conveniência, decidimos fixar a forma exponencial (EX).

Ao testar diferentes tamanhos de amostra, observamos que o desempenho do estimador na base de treinamento permaneceu semelhante em ambas as métricas de avaliação para todos os tamanhos avaliados, sendo o menor $n = 100$. Novamente, fixamos o tamanho $n = 250$ e comparamos o desempenho dos outros modelos avaliados nas mesmas amostras, concluindo não haver diferença significativa entre o SBKM e os modelos avaliados.

Ao final, analisamos o desempenho do estimador em diferentes conjuntos de treinamento com diversas taxas de censura. Repetimos o procedimento utilizado na base de crédito, mantendo o tamanho da amostra fixo em $n = 250$ e selecionando aleatoriamente falhas e censuras. As conclusões foram semelhantes: das duas métricas avaliadas, apenas o IBS foi significativamente prejudicado à medida que a porcentagem de censuras aumentou.

Entre as dificuldades enfrentadas durante o processo, a principal foi o tempo computacional necessário para otimizar os parâmetros do estimador SBKM, bem como para realizar a inferência em novas amostras. Como mencionado anteriormente, para estimar a curva de sobrevivência de um único indivíduo, é necessário conhecer a similaridade entre esse indivíduo e todos os outros da base de dados original. Esse aspecto é praticamente irrelevante para conjuntos de dados pequenos, como na base de dados CREDIT. No entanto, para a base SUPPORT, enfrentamos limitações de recursos computacionais e de tempo demandado para cada simulação.

Outro ponto relevante é que lidar com os algoritmos de otimização usados para encontrar os pesos que maximizam a função de verossimilhança não foi simples. Alguns desses

algoritmos não suportam restrições inicialmente definidas, e nem sempre seu comportamento é previsível devido a erros numéricos durante o processo de busca pela solução ótima. Apesar disso, conseguimos contornar essas dificuldades garantindo, em princípio, que as soluções convergissem para o mesmo resultado, independentemente dos valores iniciais utilizados. Investigar mais detalhadamente os algoritmos de otimização não está no escopo deste trabalho e poderia ser abordado em trabalhos futuros.

Em futuras pesquisas, pretendemos comparar o estimador de Kaplan-Meier baseado em similaridade (SBKM) com outras adaptações do estimador de Kaplan-Meier (KM) presentes na literatura, bem como com métodos de estimação por Kernel. Outra abordagem será utilizar dados de sobrevivência simulados para explorar com mais detalhes as propriedades do estimador e seu comportamento em diferentes cenários controlados. Além disso, planejamos testar diferentes coeficientes de similaridade, como a distância de Gower (Gower, 1971), e diferentes técnicas de otimização para buscar a melhor combinação de hiperparâmetros para o SBKM (Damblin *et al.*, 2013). Também temos interesse em explorar estratégias para reduzir o tempo computacional do estimador, possibilitando sua aplicação em conjuntos de dados maiores.

REFERÊNCIAS

AKAIKE, H. An approximation to the density function. **Annals of the Institute of Statistical Mathematics**, Netherlands, v. 6, n. 2, p. 127–132, 1954. Disponível em: https://doi.org/10.1007/BF02900741. Acesso em: 28 set. 2024.

ALABDALLAH, A. *et al.* The concordance index decomposition: a measure for a deeper understanding of survival prediction models. **Artificial Intelligence in Medicine**, Netherlands, v. 148, p. 102781, 2024. Disponível em: https://www.sciencedirect.com/science/article/pii/S093336572400023X. Acesso em: 28 set. 2024.

BENDA, B. B. Survival analysis of criminal recidivism of boot camp graduates using elements from general and developmental explanatory models. **International Journal of Offender Therapy and Comparative Criminology**, United States, v. 47, n. 1, p. 89–110, 2003. Disponível em: https://doi.org/10.1177/0306624X02239277. Acesso em: 28 set. 2024.

BERAN, R. Nonparametric regression with randomly censored survival data. **Technical Report**, Berkeley, 1981.

BLADT, M.; FURRER, C. Expert kaplan–meier estimation. **Scandinavian Actuarial Journal**, United Kingdom, v. 2024, n. 1, p. 1–27, 2024. Disponível em: https://doi.org/10.1080/03461238.2023.2197442. Acesso em: 28 set. 2024.

BOX-STEFFENSMEIER, J. M. *et al.* Survival analysis of faculty retention and promotion in the social sciences by gender. **PloS one**, United States, v. 10, n. 11, p. 5–32, 2015. Disponível em: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0143093. Acesso em: 28 set. 2024.

BREIMAN, L. Random forests. **Machine Learning**, Netherlands, v. 45, n. 1, p. 5–32, Oct 2001.

CHEN, G. H. Nearest neighbor and kernel survival analysis: nonasymptotic error bounds and strong consistency rates. **PLMR**, United States, v. 97, p. 1001–1010, June 2019.

CHEN, G. H. Survival kernets: scalable and interpretable deep kernel survival analysis with an accuracy guarantee. **Journal of Machine Learning Research**, United States, v. 25, n. 40, p. 1–78, 2024. Disponível em: http://jmlr.org/papers/v25/22-0667.html. Acesso em: 28 set. 2024.

CLARK, T. G. *et al.* Survival analysis part i: basic concepts and first analyses. **British Journal of Cancer**, United Kingdom, v. 89, p. 232–238, 2003. Disponível em: https://www.nature.com/articles/6601118. Acesso em: 28 set. 2024.

COLOSIMO, E. A.; GIOLO, S. R. **Análise de sobrevivência aplicada**. São Paulo: Blucher, 2006.

COX, D. R. Regression models and life-tables. **Journal of the Royal Statistical Society:** series b (methodological), United Kingdom, v. 34, n. 2, p. 187–202, 1972. Disponível em: https://doi.org/10.1111/j.2517-6161.1972.tb00899.x. Acesso em: 28 set. 2024.

DAMBLIN, G.; COUPLET, M.; IOOSS, B. Numerical studies of space filling designs: optimization of latin hypercube samples and subprojection properties. **Journal of Simulation**, United Kingdom, v. 7, p. 276–289, 2013. Disponível em: https://hal.science/hal-00848240. Acesso em: 29 set. 2024.

DAVIDSON-PILON, C. lifelines: Survival analysis in python. **Journal of Open Source Software**, United Kingdom, v. 4, n. 40, p. 1317, 2019. Disponível em: https://doi.org/10.21105/joss.01317. Acesso em: 29 set. 2024.

FERREIRA, J. C.; PATINO, C. M. O que é análise de sobrevida e quando devo utilizá-la? **Jornal Brasileiro de Pneumologia**, Brasil, v. 42, p. 77–77, 2016. Disponível em: https://doi.org/10.1590/S1806-37562016000000013. Acesso em: 29 set. 2024.

FOTSO, S. *et al.* **PySurvival: open source package for survival analysis modeling**. 2019. Disponível em: https://www.pysurvival.io/. Acesso em: 29 set. 2024.

GAYER, G. *et al.* Rule-based and case-based reasoning in housing prices. **The B.E. Journal of Theoretical Economics**, Germany, v. 7, n. 1, p. 000010220219351704 1284, 2007. Disponível em: https://doi.org/10.2202/1935-1704.1284. Acesso em: 29 set. 2024.

GILBOA, I. *et al.* Empirical similarity. **The Review of Economics and Statistics**, United States, v. 88, n. 3, p. 433–444, 2006. Disponível em: https://doi.org/10.1162/rest.88.3.433. Acesso em: 29 set. 2024.

GILBOA, I. *et al.* A similarity-based approach to prediction. **Journal of Econometrics**, Netherlands, v. 162, n. 1, p. 124–131, 2011. Disponível em: https://www.sciencedirect.com/science/article/pii/S0304407609002681. Acesso em: 29 set. 2024.

GILL, R. Glivenko-cantelli for kaplan-meier. **Mathematical Methods of Statistics**, Netherlands, v. 3, p. 76–87, 01 1994. Disponível em: https://www.researchgate.net/publication/259463065_Glivenko-Cantelli_for_Kaplan-Meier. Acesso em: 29 set. 2024.

GORDON, L.; OLSHEN, R. A. Tree-structured survival analysis. **Cancer Treatment Reports**, United States, v. 69, n. 10, p. 1065–1069, 1985.

GOURIEROUX, C.; MONFORT, A.; TROGNON, A. Pseudo maximum likelihood methods: theory. **Econometrica**, United States, v. 52, n. 3, p. 681–700, 1984. Disponível em: https://doi.org/10.2307/1913471. Acesso em: 29 set. 2024.

GOWER, J. C. A general coefficient of similarity and some of its properties. **Biometrics**, United Kingdom, v. 27, n. 4, p. 857–871, 1971. Disponível em: https://doi.org/10.2307/2528823. Acesso em: 29 set. 2024.

HOFMANN, H. **Statlog (German Credit Data) Dataset**. 1994. UCI Machine Learning Repository. Disponível em: https://doi.org/10.24432/C5NC77. Acesso em: 29 set. 2024.

HOTHORN, T. *et al.* Survival ensembles. **Biostatistics**, United Kingdom, v. 7, n. 3, p. 355–373, 2006. Disponível em: https://doi.org/10.1093/biostatistics/kxj011. Acesso em: 29 set. 2024.

HU, G.; HUFFER, F. Modified kaplan-meier estimator and nelson-aalen estimator with geographical weighting for survival data: modified kaplan-meier and nelson-aalen. **Geographical Analysis**, United States, v. 52, p. 1–21, 2019. Disponível em: https://doi.org/10.1111/gean.12185. Acesso em: 29 set. 2024.

ISHWARAN, H. *et al.* Random survival forests. **The Annals of Applied Statistics**, United States, v. 2, n. 3, p. 841 – 860, 2008. Disponível em: https://doi.org/10.1214/08-AOAS169. Acesso em: 29 set. 2024.

KALBFLEISCH, J. D.; PRENTICE, R. L. **The statistical analysis of failure time data**. New York: John Wiley and Sons, 2002.

KAPLAN, E. L.; MEIER, P. Nonparametric estimation from incomplete observations. **Journal of the American Statistical Association**, United States, v. 53, n. 282, p. 457–481, 1958. Disponível em: https://doi.org/10.2307/2281868. Acesso em: 29 set. 2024.

KASUMIGASEKI, C. ku. **The 20th Life Tables**. Japan: Statistics and Information Department, Minister's Secretariat, Ministry of Health, Labour and Welfare, Japanese Government, 2005. Disponível em: https://www.mhlw.go.jp/english/database/db-hw/lifetb20th/index.html. Acesso em: 29 set. 2024.

KNAUS, W. A. *et al.* The support prognostic model. objective estimates of survival for seriously ill hospitalized adults. study to understand prognoses and preferences for outcomes and risks of treatments. **Annals of internal medicine**, United States, v. 122, n. 3, p. 191–203, 1995. Disponível em: https://doi.org/10.7326/0003-4819-122-3-199502010-00007. Acesso em: 29 set. 2024.

KRAFT, D. **A software package for sequential quadratic programming**. Köln: Wiss. Berichtswesen d. DFVLR, 1988.

KVAMME, H.; BORGAN, Ø. The brier score under administrative censoring: problems and solutions. **Journal of Machine Learning Research**, United States, v. 24, p. 1–26, 2023.

LONGATO, E. *et al.* A practical perspective on the concordance index for the evaluation and selection of prognostic time-to-event models. **Journal of biomedical informatics**, United States, v. 108, p. 103496, 2020. Disponível em: https://doi.org/10.1016/j.jbi.2020.103496. Acesso em: 29 set. 2024.

MANTEL, N. Evaluation of survival data and two new rank order statistics arising in its consideration. **Cancer Chemotherapy Reports**, United States, v. 50, n. 3, p. 163–170, 1966.

MEHROTRA, D. V.; WEST, R. M. Survival analysis using a 5-step stratified testing and amalgamation routine (5-star) in randomized clinical trials. **Statistics in Medicine**, United States, v. 39, p. 4724–4744, 2020. Disponível em: https://doi.org/10.1002/sim.8750. Acesso em: 29 set. 2024.

PARK, S. Y. *et al.* Review of statistical methods for evaluating the performance of survival or other time-to-event prediction models (from conventional to deep learning approaches). **The Korean Society of Radiology**, Korea, v. 22, 2021. Disponível em: https://doi.org/10.3348/kjr.2021.0223. Acesso em: 29 set. 2024.

PÖLSTERL, S. scikit-survival: A library for time-to-event analysis built on top of scikit-learn. **Journal of Machine Learning Research**, United States, v. 21, n. 212, p. 1–6, 2020. Disponível em: http://jmlr.org/papers/v21/20-729.html. Acesso em: 29 set. 2024.

PRENTICE, R. L. *et al.* The analysis of failure times in the presence of competing risks. **Biometrics**, United Kingdom, v. 34, n. 4, p. 541–554, 1978. Disponível em: https://doi.org/10.2307/2530374. Acesso em: 29 set. 2024.

SALNIKOV, D. A constructive proof of the glivenko-cantelli theorem. arXiv.org, [Ithaca, N. Y.], 2021. Disponível em: https://arxiv.org/abs/2110.13236. Acesso em: 28 set. 2024. 2021.

SANCHEZ, J. D. *et al.* Prediction by empirical similarity via categorical regressors. **Machine Learning and Knowledge Extraction**, Switzerland, v. 1, n. 2, p. 641–652, 2019. Disponível em: https://doi.org/10.3390/make1020038. Acesso em: 29 set. 2024.

SANCHEZ, J. J. D. **Similaridade empírica:** o caso das variáveis explicativas categóricas. Dissertação (Mestrado) – Universidade Federal de Pernambuco, Brasil, Fevereiro 2015.

SATTEN, G. A.; DATTA, S. The kaplan-meier estimator as an inverse-probability-of-censoring weighted average. **The American Statistician**, United States, v. 55, n. 3, p. 207–210, 2001. Disponível em: https://doi.org/10.1198/000313001317098185. Acesso em: 29 set. 2024.

SIMON, N. *et al.* Regularization paths for cox's proportional hazards model via coordinate descent. **Journal of Statistical Software**, United States, v. 39, n. 5, p. 1–13, 2011. Disponível em: https://doi.org/10.18637/jss.v039.i05. Acesso em: 29 set. 2024.

SONG, Y.-Y.; LU, Y. Decision tree methods: applications for classification and prediction. **Shanghai Arch Psychiatry**, China, v. 27, n. 2, p. 130–135, 2015. Disponível em: https://doi.org/10.11919%2Fj.issn.1002-0829.215044. Acesso em: 29 set. 2024.

STRAPASSON, E. **Comparação de modelos com censura intervalar em análise de sobrevivência**. Tese (Doutorado) – Escola Superior de Agricultura Luiz de Queiroz, Universidade de São Paulo, Piracicaba, 2007.

UNO, H. *et al.* On the c-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. **Statistics in medicine**, United States, v. 30, p. 1105–17, 2011.

VEZHNEVETS, A.; BARINOVA, O. Avoiding boosting overfitting by removing confusing samples. In: *In:* MACHINE Learning: ECML 2007. Berlin, Heidelberg: Springer Berlin Heidelberg, 9007. p. 430–441. Disponível em: https://doi.org/10.1007/978-3-540-74958-5_40. Acesso em: 29 set. 2024.

WANG, P. *et al.* Machine learning for survival analysis: a survey. **Association for Computing Machinery**, United States, v. 51, n. 6, 2019. Disponível em: https://doi.org/10.1145/3214306. Acesso em: 29 set. 2024.

WANG, S.; GITTENS, A.; MAHONEY, M. W. Scalable kernel k-means clustering with nyström approximation: relative-error bounds. **Journal of Machine Learning Research**, United States, v. 20, n. 1, p. 431–479, 2019.

WEGNER, P. A technique for counting ones in a binary computer. **Association for Computing Machinery**, United States, v. 3, n. 5, p. 322, 1960. Disponível em: https://doi.org/10.1145/367236.367286. Acesso em: 29 set. 2024.

WEHRENS, R. *et al.* The bootstrap: a tutorial. **Chemometrics and Intelligent Laboratory Systems**, Netherlands, v. 54, n. 1, p. 35–52, 2000. Disponível em: https://doi.org/10.1016/S0169-7439(00)00102-7. Acesso em: 29 set. 2024.

WILLEMS, S. J. **Inverse probability censoring weights for routine outcome monitoring data**. Dissertação (Mestrado) – Universiteit Leiden, Netherlands, June 2014.

WILSON, D.; MARTINEZ, T. Reduction techniques for instance-based learning algorithms. **Machine Learning**, Netherlands, v. 38, p. 257–286, 01 2000. Disponível em: https://doi.org/10.1023/A:1007626913721. Acesso em: 29 set. 2024.

XU, Y.; GOODACRE, R. On splitting training and validation set: a comparative study of cross-validation, bootstrap and systematic sampling for estimating the generalization performance of supervised learning. **Journal of Analysis and Testing**, Singapore, v. 2, n. 3, p. 249–262, 2018. Disponível em: https://doi.org/10.1007/s41664-018-0068-2. Acesso em: 29 set. 2024.

YANG, Z. *et al.* Prognostic modeling of predictive maintenance with survival analysis for mobile work equipment. **Scientific Reports**, United Kingdom, v. 12, n. 1, p. 8529, 2022. Disponível em: https://doi.org/10.1038/s41598-022-12572-z. Acesso em: 29 set. 2024.

ZHANG, Z. Variable selection with stepwise and best subset approaches. **Annals of translational medicine**, China, v. 4, n. 7, p. 136, 2016. Disponível em: https://doi.org/10.21037/atm.2016.03.35. Acesso em: 29 set. 2024.

ZHOU, M. **Empirical likelihood method in survival analysis**. Florida: CRC Press, 2019.

APÊNDICE A – DETALHES DE OTIMIZAÇÃO

Neste estudo, para otimizar os hiperparâmetros dos modelos de referência usados para comparação, empregamos a técnica conhecida como *Grid Search*. Essa abordagem consiste em construir uma "grade" com todos os possíveis valores de hiperparâmetros especificados e, em seguida, avaliar o desempenho do modelo em uma base de validação para cada combinação de hiperparâmetros.

Listamos agora as grades de hiperparâmetros utilizadas para os diferentes métodos:

Cox Proportional-Hazards (COX) usando a biblioteca *lifelines* e a classe `CoxPHFitter`:

- `l1_ratio`: 0, 0.01, 0.05, 0.5, 1
- `penalizer`: 0, 0.01, 0.05, 0.5, 1

Elastic-Net Cox (EN-COX) usando a biblioteca *scikit-survival* e a classe `CoxnetSurvivalAnalysis`:

- `l1_ratio`: 0, 0.01, 0.05, 0.5, 1
- `n_alphas`: 1, 10, 50, 100
- `normalize`: `True`, `False`

Distribuição de Weibull (WEIBULL) usando a biblioteca *lifelines* e a classe `WeibullAFTFitter`:

- `l1_ratio`: 0, 0.01, 0.05, 0.5, 1
- `penalizer`: 0, 0.01, 0.05, 0.5, 1
- `model_ancillary`: `True`, `False`

Survival Tree (ST) usando a biblioteca *scikit-survival* e a classe `SurvivalTree`:

- `splitter`: "best", "random"
- `max_depth`: 5, 10, 50
- `min_samples_split`: 2, 4, 6, 8, 10
- `min_samples_leaf`: 3, 6, 9
- `max_leaf_nodes`: 5, 10, 50

Random Survival Forest (RSF) usando a biblioteca *scikit-survival* e a classe `RandomSurvivalForest`:

- `n_estimators`: 5, 10, 50, 100
- `max_depth`: 5, 10, 50
- `min_samples_split`: 2, 4, 6, 8, 10
- `min_samples_leaf`: 3, 6, 9
- `max_leaf_nodes`: 5, 10, 50

Gradient-Boosted Cox (GB-COX) usando a biblioteca *scikit-survival* e a classe `GradientBoostingSurvivalAnalysis`:

- `learning_rate`: 0.01, 0.1, 1, 10
- `n_estimators`: 5, 10, 50, 100
- `max_depth`: 5, 10, 50
- `min_samples_split`: 2, 4, 6, 8, 10
- `min_samples_leaf`: 3, 6, 9
- `max_leaf_nodes`: 5, 10, 50