Universidade Estadual Paulista “Júlio de Mesquita Filho”
Instituto de Biociências – Câmpus de Botucatu
Programa de Pós-graduação em Biometria

Modelos de predição para dados censurados aplicados a pacientes com doença renal crônica em terapia de substituição renal

Agda Jéssica de Freitas Galletti

Botucatu
2022

Agda Jéssica de Freitas Galletti

Modelos de predição para dados censurados aplicados a pacientes com doença renal crônica em terapia de substituição renal

Tese de Doutorado apresentada ao Curso do Programa de Pós-graduação em Biometria da Universidade Estadual Paulista "Júlio de Mesquita Filho" como parte dos requisitos necessários para a obtenção do título de Doutora em Biometria.

Orientadora: Profa. Dra. Liciana Vaz de Arruda Silveira

Coorientadora: Profa. Dra. Daniela Ponce

Botucatu
2022

FICHA CATALOGRÁFICA ELABORADA PELA SEÇÃO TÉC. AQUIS. TRATAMENTO DA INFORM.
DIVISÃO TÉCNICA DE BIBLIOTECA E DOCUMENTAÇÃO - CÂMPUS DE BOTUCATU - UNESP

BIBLIOTECÁRIA RESPONSÁVEL: ROSANGELA APARECIDA LOBO-CRB 8/7500

Galletti, Agda Jéssica Freitas.
Modelos de predição para dados censurados aplicados a pacientes com doença renal crônica em terapia de substituição renal / Agda Jéssica Freitas Galletti. - Botucatu, 2022

Tese (doutorado) - Universidade Estadual Paulista "Júlio de Mesquita Filho", Instituto de Biociências de Botucatu
Orientador: Liciana Vaz de Arruda Silveira
Coorientador: Daniela Ponce
Capes: 90194000

1. Análise de sobrevivência (Biometria). 2. Aprendizado do computador. 3. Diálise peritoneal. 4. Hemodiálise. 5. Predição (Logica).

Palavras-chave: Análise de sobrevivência; Aprendizado de máquina; Diálise peritoneal; Hemodiálise; Modelos de predição.

unesp

UNIVERSIDADE ESTADUAL PAULISTA

Câmpus de Botucatu

ATA DA DEFESA PÚBLICA DA TESE DE DOUTORADO DE AGDA JÉSSICA DE FREITAS GALLETTI, DISCENTE DO PROGRAMA DE PÓS-GRADUAÇÃO EM BIOMETRIA, DO INSTITUTO DE BIOCIÊNCIAS - CÂMPUS DE BOTUCATU.

Aos 17 dias do mês de agosto do ano de 2022, às 08:00 horas, por meio de Videoconferência, realizou-se a defesa de TESE DE DOUTORADO de AGDA JÉSSICA DE FREITAS GALLETTI, intitulada **Modelos de predição para dados censurados aplicados a pacientes com doença renal crônica em terapia de substituição renal**. A Comissão Examinadora foi constituída pelos seguintes membros: Profa. Dra. LICIANA VAZ DE ARRUDA SILVEIRA (Orientador(a) - Participação Virtual) do(a) Departamento de Bioestatística, Biologia Vegetal, Parasitologia e Zoologia / Instituto de Biociências de Botucatu - UNESP , Prof. Dr. THIAGO SANTOS MOTA (Participação Virtual) do(a) FATEC / Faculdade de Tecnologia de Botucatu, Prof. Dr. MAICON VINICIUS GALDINO (Participação Virtual) do(a) SEDUC / Secretaria da Educação, Profa. Dra. SILVIA HELENA MODENESE GORLA DA SILVA (Participação Virtual) do(a) Coordenadoria de Engenharia Agronômica / UNESP/Registro, Prof. Dr. ROGERIO ANTONIO DE OLIVEIRA (Participação Virtual) do(a) Departamento de Bioestatística, Biologia Vegetal, Parasitologia e Zoologia / Instituto de Biociências de Botucatu - UNESP. Após a exposição pela doutoranda e arguição pelos membros da Comissão Examinadora que participaram do ato, de forma presencial e/ou virtual, a discente recebeu o conceito final:_____ APROVADA _____. Nada mais havendo, foi lavrada a presente ata, que após lida e aprovada, foi assinada pelo(a) Presidente(a) da Comissão Examinadora.

Profa. Dra. LICIANA VAZ DE ARRUDA SILVEIRA

À minha amada mãe e aos meus amigos e professores dedico este trabalho. A presença de vocês durante esta incrível e louca jornada tornou tudo mais fácil. Sinto-me grata!
Dedico também a mim por perseverar e persistir ao longo de anos longe de casa na bela e charmosa cidade de Botucatu.

Agradecimentos

Agradeço a Deus e a todos que de alguma forma contribuíram para a realização deste trabalho, em especial:

À minha querida orientadora Profa. Dra. Liciana Vaz por todo suporte e orientação. Além disso, pela amizade, parceria, incansáveis incentivos e conselhos, pelas ótimas conversas e pelas oportunidades de conhecimento a mim ofertadas.

À minha coorientadora Profa. Dra. Daniela Ponce por ter topado embarcar nesta jornada, por todo suporte, parceria e conhecimento compartilhado.

À Unidade de Diálise do Hospital das Clínicas da Faculdade de Medicina de Botucatu e seus pacientes que permitiram o acesso aos prontuários médicos e por todo suporte oferecido.

A todos professores e funcionários do Departamento de Bioestatística, Biologia Vegetal, Parasitologia e Zoologia, Unesp, Botucatu, pelo apoio, gentileza e disponibilidade.

Ao programa de Pós-graduação em Biometria e à Unesp pelas diversas oportunidades a mim oferecidas em docência, por meio de bolsas de estágio à docência e participação dos Programa de Atividades e Aperfeiçoamento em Docência no Ensino Superior (PAADES) e Programa Formação Didático-Pedagógica para Cursos de Modalidade a Distância, que permitiram me tornar uma profissional melhor e ainda contribuíram com a continuidade dos meus estudos.

À minha família que sempre esteve ao meu lado, principalmente à minha mãe por todo apoio, amor e confiança.

A todos os alunos de mestrado e doutorado do programa de pós graduação em Biometria, que direta ou indiretamente, contribuíram positivamente para que este trabalho fosse realizado.

Aos meus amigos que fizeram desta jornada mais agradável e feliz. Obrigada pelos abraços, sorrisos, eventuais momentos de lazer, sessões de cinema, dias e noites de estudos, alimentações saudáveis, besteiras também, festinhas, companheirismo, confiança e inúmeros incentivos. Destacando as minhas albercats e companheiras de casa por serem luz em meio ao caos que foi esta pandemia.

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - CAPES pelo incentivo a pós-graduações de todos país e por oportunizar a formação de pesquisadores e popularização do conhecimento científico.

Resumo

A doença renal crônica, caracterizada pela alteração da função renal de modo progressivo e irreversível, é não transmissível e uma das principais causas de mortalidade em todo o mundo. Pacientes em estágio muito avançado da doença necessitam iniciar a terapia renal substitutiva, isto é, diálise peritoneal (DP), hemodiálise (HD) ou transplante renal. Segundo a Sociedade Brasileira de Nefrologia, em 2020, 144.779 pacientes estavam em diálise, sendo que, anualmente, cerca de 20 mil novos pacientes iniciaram o tratamento de hemodiálise e apresentam taxa de mortalidade de 15% ao ano. A análise estatística pode auxiliar no planejamento e no desenvolvimento de estratégias para o controle da doença, importante para a manutenção, gestão, prevenção e avaliação de políticas voltadas à saúde. Além disso, a predição do prognóstico de doenças é um dos principais objetivos para médicos e gestores de saúde pública, que pode ser utilizada para direcionamento de intervenções preventivas e podem fornecer a probabilidade do paciente ter ou desenvolver uma determinada doença. Portanto, este trabalho tem como objetivo obter os fatores associados à mortalidade de pacientes em tratamento dialítico por meio de algoritmos de aprendizado de máquina com inclusão dos tempos de sobrevivência censurados. Como resultado, o modelo de Cox teve melhor desempenho preditivo. E também, pacientes que iniciaram a terapia renal substitutiva em hemodiálise têm maior risco de morte do que aqueles que iniciaram com diálise peritoneal. Entretanto, pacientes internados e que iniciaram o tratamento em HD terão menor risco de morte do que aqueles que iniciaram com DP. Além disso, independente do tratamento iniciado, pacientes que foram internados têm maior risco de morte. O modelo de Cox ajustado indicou a idade como um fator de risco, conforme indicado na literatura. Diferentemente, a presença de doenças de base não foram significativas para explicar o tempo de vida dos pacientes em tratamento.

Palavras-chave: aprendizado de máquina; diálise peritoneal; hemodiálise; modelo de predição de sobrevivência.

Abstract

Chronic kidney disease, characterized by progressive and irreversible loss of kidney function, is noncommunicable and one of the main causes of mortality worldwide. Patients in an advanced stage of the disease need to start renal replacement therapy, that is, peritoneal dialysis (PD), hemodialysis (HD) or kidney transplant. According to the Brazilian Society of Nephrology, in 2020, 144,779 patients were on dialysis, with around 20,000 new patients starting hemodialysis annually and a mortality rate of 15% per year. Statistical analysis can assist in the planning and development of strategies for disease control, which is important for the maintenance, management, prevention and evaluation of health policies. In addition, the prediction of disease prognosis is one of the main goals for physicians and public health managers, which can be used to target preventive interventions and can provide the probability of the patient having or developing a certain disease. Therefore, this work aims to obtain the factors associated with mortality of patients on dialysis through machine learning algorithms including censored survival times. As a result, the Cox model had the best predictive performance. Also, patients who started renal replacement therapy on hemodialysis have a higher risk of death than those who started on peritoneal dialysis. However, a patient who is hospitalized and started on HD will have a lower risk of death than a patient who started on PD. In addition, regardless of the treatment initiated, patients who are hospitalized have a higher risk of death. The adjusted Cox model indicated age as a risk factor, according to some references in the literature. Unlike, the presence of underlying diseases was non-significant to explain the lifetime of patients under treatment.

Keywords: machine learning; peritoneal dialysis; hemodialysis; survival prediction model.

Lista de figuras

Lista de tabelas

Lista de abreviaturas e siglas

AIC	Critério de Informação de Akaike.
AM	Aprendizado de máquima.
AS	Análise de sobrevivência.
CART	*Classification and Regression Tree* (Árvores de Classificação e Regressão).
CVC	Cateter venoso central.
DP	Diálise peritoneal.
DCNT	Doenças crônicas não transmissíveis.
DCR	Doenças renal crônica.
DRET	Doença renal crônica estágio terminal.
EKM	Estimador de Kaplan-Meier.
FAV	Fístula arteriovenosa.
HD	Hemodiálise.
MVS	Máquinas de vetores de suporte.
pmp	Por milhão da população.
RNA	Redes neurais artificiais.
RVS	Regressão do vetor de suporte.
SBN	Sociedade Brasileira de Nefrologia.
TFG	Taxa de filtração glomerular.
TRS	Terapia renal substitutiva.
TSR	Terapia de substituição renal.

Lista de Códigos em R

Sumário

1 INTRODUÇÃO

As doenças crônicas não transmissíveis (DCNT) são as principais causas de mortalidade e incapacitação em todo o mundo (SILVA et al., 2017). O Instituto Brasileiro de Geografia e Estatística – IBGE (2014) destaca que as DCNT constituem o problema de saúde de maior magnitude e correspondem por mais de 70% das causas de mortes no Brasil. Segundo o secretário de Vigilância em Saúde do Ministério da Saúde (SVS/MS), Arnaldo Medeiros, durante a Semana das Doenças Crônicas não Transmissíveis, destacou que as DCNT matam cerca de 41 milhões de pessoas a cada ano, isto é, cerca de 71% de todas as mortes no mundo. Sendo que, 77% dessas mortes ocorrem em países de baixa e média renda (BRASIL, 2021b).

Dentre as DCNT, tem-se a doença renal crônica (DRC), caracterizada pela diminuição da função renal, que em muitas vezes tem evolução assintomática. O estágio avançado da doença exige que pacientes necessitem de terapia renal substitutiva (TRS) como a diálise contínua (hemodiálise e diálise peritoneal) ou o transplante de rim (MARINHO et al., 2017). De acordo com Marcelo Mazza, presidente da Sociedade Brasileira de Nefrologia (SBN), a DRC pode ser considerada epidêmica, ao passo que, atinge um a cada dez adultos e a taxa de incidência aumenta cada vez mais.

Anualmente, no Brasil, cerca de 20 mil novos pacientes iniciam o tratamento de hemodiálise e taxa de mortalidade de 15% (BRASILIA, 2020). Segundo o Censo Brasileiro de Diálise 2020, financiado pela SBN, “144.779 pacientes estão em diálise e as taxas estimadas de prevalência e incidência de pacientes por milhão da população (pmp) foram 684 e 209, respectivamente. Dentre os pacientes prevalentes, 92,6% estavam em hemodiálise (HD) e 7,4% em diálise peritoneal (DP)” (NERBASS et al., 2022). Diante da alta prevalência da DRC, a qual cursa com elevada mortalidade, é importante implementar medidas com foco no fortalecimento da atenção primária, conforme indicado pelo secretário de Vigilância em Saúde, Arnaldo Medeiros (BRASIL, 2021b). Estudos relacionados ao tema contribuem para o planejamento, gestão e avaliação de políticas voltadas à saúde e assistência social de pacientes com doença renal crônica sendo a análise estatística desses dados ferramenta fundamental.

O tratamento de algumas enfermidades requer o acompanhamento do paciente até a cura ou o óbito, sendo natural que o acompanhamento de um ou outro paciente seja interrompido, causando uma observação parcial da resposta. Por consequência, a análise de sobrevivência é uma metodologia muito utilizada em pesquisas médicas desta natureza (COLLETT, 2003;

KLEINBAUM; KLEIN, 2012). Além disso, a predição do prognóstico de doenças é um dos objetivos principais para médicos e gestores de saúde pública, que pode ser utilizada para direcionamento de intervenções preventivas, estabelecendo um critério sobre a triagem, tratamento em grupos de alto risco, investigação diagnóstica e escolha de terapia, a fim de evitar o óbito dos pacientes (STEYERBERG, 2009; CHEN, 2020).

A principal característica deste tipo de modelo é a criação de um algoritmo preditivo com boa precisão, consequentemente, ele não precisa ser o correto, ou seja, a análise de resíduos ou da qualidade do ajuste são dispensados (HASTIE; TIBSHIRANI; FRIEDMAN, 2017). Metodologias como a aprendizado de máquina (AM), mecanismo para a busca de padrões e a construção de inteligência em uma máquina para poder aprender, o que implica que poderá ser melhor no futuro a partir de sua própria experiência (GOLLAPUDI, 2016), são ferramentas utilizadas para predição. No processo de aprendizado é levado em conta o tipo de variável a ser estudada, isto é, quando a variável de interesse é quantitativa, tem-se problemas de regressão, quando qualitativa, classificação. É possível encontrar diversos trabalhos como de Maccariello et al. (2008), Segal et al. (2020), Kim et al. (2021) e Monaghan et al. (2021) que optaram em obter modelos preditivos para pacientes renais, entretanto, focados em objetivos distintos e centrados em modelos de classificação ou regressão. No entanto, abordagens para dados de sobrevivência, como serão apresentados, ainda são poucas e seus desenvolvimentos recentes (WANG; LI; REDDY, 2019). Além disso, Spooner et al. (2020) observaram que os resultados obtidos via aprendizado de máquina podem fornecer alternativas mais precisas aos métodos tradicionais de análise de sobrevivência, como o modelo de riscos proporcionais de Cox, principalmente, na presença de dados de alta dimensão. Por estes motivos, o desenvolvimento de modelos de predição é uma interessante abordagem no estudo dos tempos de sobrevivência de pacientes com doença renal crônica.

O objetivo geral da pesquisa foi desenvolver um estudo comparativo entre os modelos de aprendizado de máquina para dados censurados a fim de identificar os fatores de risco ou proteção que interferem na sobrevivência da população de pacientes que apresentam doença renal crônica em tratamento dialítico. Além de apresentar de forma didática como utilizar o *software* R para treinar modelos de aprendizado de máquina de sobrevivência.

Referente a estrutura desta tese, a organização está disposta em seis capítulos. O segundo capítulo apresenta-se um resumo sobre a DRC; no terceiro e quarto capítulos, os conceitos básicos da análise de sobrevivência e algoritmos de aprendizado de máquina com a inclusão da censura. Também será dedicado a apresentação de alguns hiperparâmetros dos modelos de predição e o passo a passo da análise utilizando o *software* R; o quinto capítulo, compõem a análise estatística do conjunto de dados e a descrição dos resultados; e por fim, apresenta-se a discussão e as considerações finais, seguido pelas referências bibliográficas e anexos.

2 DOENÇA RENAL CRÔNICA

O ser humano, naturalmente, possui dois rins, que medem aproximadamente 12 centímetros, pesam por volta de 150 gramas e estão localizados atrás das últimas costelas. Os rins possuem diversas funções no organismo, dentre elas: a eliminação de toxinas do sangue, o controle hidroeletrolítico e metabólico, bem como, a regulação da pressão sanguínea e da formação do sangue e ossos. Clinicamente, a forma mais eficaz para avaliar as funções renais é a excretora por ter maior correlação com os desfechos clínicos (BRASIL, 2014; SBN, 2022a).

A DRC é caracterizada pela perda gradual, progressiva e irreversível da função dos rins. Assim, considera-se que qualquer indivíduo porta DRC, independente da causa, se apresentar por pelo menos três meses consecutivos uma taxa de filtração glomerular (TFG) $< 60ml/min/1,73m^2$. Entretanto, nas situações em que a $TFG \geq 60ml/mim/1,73m^2$, é considerado portador de DRC o paciente que tiver pelo menos um marcador de dano renal parenquimatoso ou alteração no exame de imagem (BRASIL, 2014; MARINHO et al., 2017).

Atualmente, a DRC é classificada em 5 estágios funcionais de acordo com a TFG, como mostrado na Tabela 1. Nela, apresenta-se também a proteinúria (ou albuminúria), marcador de dano renal comumente utilizado para classificação dos pacientes. Entretanto, existem outros marcadores de dano renal que podem ser analisados, por exemplo, alterações na urina (hematúria glomerular), em imagens ultrassonográficas (rins policísticos) ou alterações histopatológicas vistas em biópsias renais (BASTOS; KIRSZTAJN, 2011).

Tabela 1 – Estadiamento da doença renal crônica proposto pelo KDOQI e atualizado pelo *National Collaborating Centre for Chronic Condition*

Estágios da DRC	**Taxa de filtração glomerular***	**Proteinúria**
1	≥ 90	Presente
2	60-89	Presente
3A	45-59	Presente ou ausente
3B	30-44	Presente ou ausente
4	15-29	Presente ou ausente
5	<15	Presente ou ausente

$*mL/min/1,73m^2$
KDOQI: Kidney Disease Outcomes Quality Initiative
Fonte: Bastos e Kirsztajn (2011)

Os pacientes com DRC são identificados de formas distintas: detectados na população

geral e aqueles que são encaminhados para centros de nefrologia. Em geral, a DRC da comunidade é caracterizada pela presença de idosos, com comorbidades, TFG decrescendo de forma lenta e com a maioria morrendo antes de atingir o estágio avançado da DRC. Enquanto que, os com DRC encaminhada têm nefropatia significativa, herdada ou adquirida, antecipando a perda da função renal e a evolução para DRC-5D (JOHNSON; FEEHALLY; FLOEGE, 2016).

Alguns estudos indicam que a idade (idosos), raça ou etnicidade (não caucasianos), genética, peso ao nascer (baixo), hipertensão arterial sistêmica, diabetes melito, doença cardiovascular, albuminúria, obesidade ou síndrome metabólica, dislipidemia, hiperuricemia, tabagismo, baixo nível socioeconômico e exposição a nefrotóxicos (anti-inflamatórios não esteroides (AINEs), analgésicos, fitoterápicos populares, metais pesados - como o chumbo) são fatores de risco modificáveis e não modificáveis associados a etilogia e progressão da DRC (JOHNSON; FEEHALLY; FLOEGE, 2016).

Sabe-se que nos estágios iniciais da DRC os pacientes, muitas vezes, são assintomáticos, por isso, investigações rotineiras são importantes, especialmente, naqueles pacientes com fatores de risco médico ou sociodemográfico para DRC. Desta forma, a dosagem de creatinina no sangue e o exame de urina de rotina são exames simples e acessíveis, que podem ser incluídos no *check up*, para obter o diagnóstico da doença. Por exemplo, a presença de proteína ou albumina, assim como, de sangue no exame de urina são alertas particularmente importantes para possível doença dos rins (BASTOS; KIRSZTAJN, 2011; SBN, 2022b).

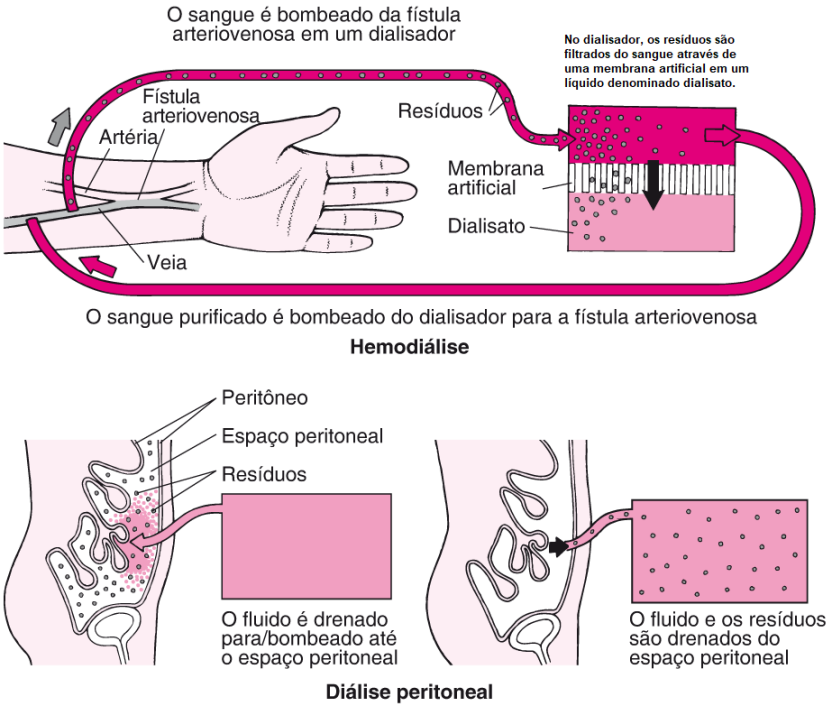
Bastos e Kirsztajn (2011) indicam três pilares para o tratamento da DRC: o diagnóstico precoce da doença, o encaminhamento imediato para tratamento nefrológico e a implementação de medidas para preservar a função renal.

Quando o paciente DRC encontra-se no estágio 5, também conhecida como insuficiência renal crônica ou doença renal crônica estágio terminal (DRET), há a necessidade de iniciar a terapia renal substitutiva (TRS). Os tratamentos disponíveis para os nefrologistas são a diálise peritoneal (DP), hemodiálise (HD), a mais frequente, e o transplante renal (BRASIL, 2014; DIAS et al., 2016).

O tratamento hemodiálise é realizado por meio de uma máquina e um dialisador que limpa e filtra o sangue. A máquina recebe o sangue do paciente por um acesso vascular por meio de um cateter ou uma fístula arteriovenosa e, em seguida, é impulsionado por uma bomba até o dialisador para ser exposto a uma solução de diálise (dialisato). Por fim, uma membrana semipermeável, responsável pelo processo de filtração, retira as toxinas em excesso e o sangue é tratado e devolvido ao paciente pelo acesso vascular. Enquanto que, a diálise peritoneal ocorre dentro da cavidade abdominal, acessada por meio de um cateter flexível, com auxílio do peritônio e da infusão da solução de diálise. Nesta abordagem, o peritônio, membrana porosa e semipermeável, que reveste os órgãos abdominais, funciona como um filtro natural. Ambos tratamentos têm a finalidade de remover os resíduos do sangue, permitindo o controle da pressão arterial, da concentração dos níveis do sódio, potássio, uréia e creatinina (BRASIL, 2021a; SBN, 2021a; SBN, 2021b).

Na Figura 1 é possível entender melhor as diferenças entre os tratamentos. A Figura 1 mostra, esquematicamente, as principais características técnicas de cada terapia.

Figura 1 – Comparação da hemodiálise com a diálise peritoneal



Fonte: Hechanova (2020)

Os pacientes com DRC 5D são aptos ao tratamento tanto com DP quanto HD. Estudo comparativos falharam em indicar uma vantagem de sobrevida consistente com qualquer das modalidades, ou seja, não há evidência de superioridade de um método em relação ao outro no que diz respeito à mortalidade geral dentro dos dois primeiros anos de tratamento (JOHNSON; FEEHALLY; FLOEGE, 2016; MENDES et al., 2017).

A diálise peritoneal, embora historicamente tenha sido muito utilizada na nefrologia, por razões não totalmente claras tem sido subutilizada com o decorrer dos anos, sobretudo em pacientes incidentes em TRS. São consideradas razões para a subutilização da DP como modalidade de TRS: a "percepção" de que é inferior à HD, devido ao fato da HD estar associada à maior tecnologia; as complicações infecciosas, mecânicas e metabólicas associadas ao método; o melhor reembolso financeiro com a HD e as dificuldades com o implante do cateter peritoneal (PONCE; BRABOA; BALBIA, 2018).

Dias et al. (2017) sugerem que "a modalidade de DP pode ser uma alternativa viável e segura para a HD, não apenas em casos planejados, mas também no cenário não planejado". A diálise não planejada pode ser definida como o início da HD sem um acesso vascular definitivo funcional, ou seja, usando um cateter venoso central (CVC) ou como início da DP em menos de 7 dias após a sua implantação (Ivarsen e Povlsen (2014) *apud* Dias et al. (2016)).

A maioria dos pacientes com DRC 5D inicia a TRS de maneira não planejada. Ivarsen e

Povlsen (2014) avaliaram retrospectivamente o Registro Dinarmaquês de Nefrologia, no período de 2008 a 2011, e verificaram que 50% dos pacientes incidentes em TRS o fizeram de maneira não planejada. Tal prática não é considerada ideal pois o CVC está associado ao aumento de hospitalização, da mortalidade e a altas taxas de bacteremia desses pacientes quando comparados ao uso de uma fístula arteriovenosa (FAV), enxerto ou DP (CHAUDHARY; SANGHA; KHANNA, 2011; DIAS et al., 2016; DIAS et al., 2017; MENDES et al., 2017).

A peritonite é a inflamação da serosa que recobre as paredes internas e as vísceras abdominais e é a principal causa de falência de técnica (GONZáLEZ et al., 2015; PONCE; BRABOA; BALBIA, 2018) e a transferência para a HD. A detecção precoce e o tratamento adequado da peritonite resultam em potenciais benefícios para qualidade de vida, longevidade e redução de custos associados ao cuidado em saúde (MARINHO et al., 2017).

Com melhores técnicas de implante de cateter e métodos automatizados, a incidência de peritonite foi reduzida e o risco de infecção na DP é semelhante a outras formas de purificação sanguínea extracorpórea (PONCE; BRABOA; BALBIA, 2018). Além disso, tem sido observado uma melhora na sobrevida do paciente e da técnica ao longo dos anos.

3 ANÁLISE DE SOBREVIVÊNCIA

A análise de sobrevivência (AS) é um conjunto de técnicas para análise de dados em que se deseja observar o tempo até a ocorrência de um evento de interesse, definido como tempo de falha (KLEINBAUM; KLEIN, 2012). A principal característica desta metodologia é a presença de censura.

Frequentemente, para os indivíduos em estudo nem sempre ocorre o evento de interesse, ou a informação do indivíduo se perde durante o seguimento, resultando em observações parciais ou incompletas. Essa ocorrência é chamada de censura (COLLETT, 2003; COLOSIMO; GIOLO, 2006; KLEINBAUM; KLEIN, 2012).

O tempo de falha do indivíduo t_i, $i = 1, ..., n$, assume o menor valor entre t_i^*, tempo de sobrevivência até a falha, e u_i, tempo de falha, ou seja,

$$t_i = min(t_i^*, u_i), \quad i = 1, ..., n,$$

enquanto que, a variável indicadora de falha (censura) é definida como,

$$\delta_i = \begin{cases} 0, & \text{se } t_i \text{ é um tempo de censura} \\ 1, & \text{se } t_i \text{ é um tempo de falha} \end{cases}.$$

A função de sobrevivência é uma das principais características avaliadas na área de AS e refere-se a probabilidade de uma observação não falhar até um certo tempo t, ou seja, a probabilidade de uma observação sobreviver ao tempo t (COLOSIMO; GIOLO, 2006). Em outros termos,

$$S(t) = P(T \geq t) = 1 - F(t), t > 0,$$

satisfazendo às seguintes propriedades:

i) $S(0) = 1$;
ii) $\lim_{t \to \infty} S(t) = 0$; e
iii) $S(t)$ é decrescente.

Em contrapartida, "a função de risco é utilizada para expressar o risco (instantâneo) de morte em algum momento t e pode ser obtida a partir da probabilidade de um indivíduo morrer no tempo t, condicional ao fato de ele ter sobrevivido até aquele momento" (COLLETT, 2003).

Sendo assim, denotada por:

$$h(t) = \lim_{\Delta t \to 0} \left(\frac{P(t \leqslant T < t + \Delta t | T \geq t)}{\Delta t} \right) = \frac{f(t)}{S(t)}.$$

Uma técnica não paramétrica utilizada para se estimar a curva de sobrevivência na presença de censura é o estimador de Kaplan-Meier (EKM), também denominado de estimador produto-limite. O EKM é um estimador de máxima verossimilhança para $S(t)$, fracamente consistente, converge assintoticamente para um processo gaussiano e definido por (KAPLAN; MEIER, 1958; COLOSIMO; GIOLO, 2006):

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

tal que:

- $t_j, j = 1, ...k$, representa o j-ésimo tempo de falha. Os tempos devem ser ordenados e distintos;
- d_j é o número de falhas em t_j;
- n_j é o número de indivíduos em risco até t_j, ou seja, indivíduos que não falharam e não foram censurados até o instante imediante anterior a t_j.

A partir das propriedades do estimador de máxima verossimilhança de $S(t)$, utilizando a fórmula de Greenwood é possível obter a variância assintótica do EKM, expressa por:

$$\widehat{Var}(\hat{S}(t)) = [\hat{S}(t)]^2 \sum_{j:t_j<t} \frac{d_j}{n_j(n_j - d_j)}. \tag{3.1}$$

O intervalo aproximado de $100(1-\alpha)\%$ de confiança para $S(t)$ é expresso por:

$$\hat{S}(t) \pm z_{\alpha/2} \sqrt{\widehat{Var}\left(\hat{S}(t)\right)},$$

em que, $z_{\alpha/2}$ é o percentil $(1-\alpha/2)100\%$ da distribuição normal padrão. Nas situações em que $S(t)$ assume valores extremos alguns intervalos estimados podem não estar contidos em $[0, 1]$. Desta forma, é necessário transformar $\hat{S}(t)$, como por exemplo, por meio:
i) transformação logística: $\log\{S(t)/(1-S(t)\}$ ou
ii) transformação complemento log-log: $log\{-logS(t)\}$.

Considerando a transformação, $\hat{U}(t) = log\{-log\hat{S}(t)\}$, tem-se que a variância assintótica é expressa por (COLOSIMO; GIOLO, 2006):

$$\widehat{Var}\left(\hat{U}(t)\right) = \frac{\sum_{j:t_j<t} \frac{d_j}{n_j(n_j - d_j)}}{\left[log\hat{S}(t)\right]^2}. \tag{3.2}$$

O intervalo aproximado de $100(1-\alpha)\%$ de confiança para $S(t)$, pode ser obtido a partir do intervalo aproximado de $U(t)$, conforme demonstrado em Galletti (2018):

$$\begin{aligned} IC_{(1-\alpha)100\%} &= \hat{U}(t) \pm z_{\alpha/2}\sqrt{\widehat{Var}(\hat{U}(t))} \\ &= log\{-log\hat{S}(t)\} \pm z_{\alpha/2}\sqrt{\widehat{Var}(\hat{U}(t))} \\ &= -\exp\left\{log\{-log\hat{S}(t)\} \pm z_{\alpha/2}\sqrt{\widehat{Var}(\hat{U}(t))}\right\} \\ &= log\hat{S}(t)\exp\left\{\pm z_{\alpha/2}\sqrt{\widehat{Var}(\hat{U}(t))}\right\} \\ &= \exp\left\{log\hat{S}(t)\exp\left\{\pm z_{\alpha/2}\sqrt{\widehat{Var}(\hat{U}(t))}\right\}\right\} \\ &= [\hat{S}(t)]^{\exp\left\{\pm z_{\alpha/2}\sqrt{\widehat{Var}(\hat{U}(t))}\right\}} \end{aligned}$$

Kalbfleisch e Prentice (2002) e Collett (2003) discutem detalhadamente os passos da demonstração da Equação 3.1 e também da transformação de $S(t)$ que resultou na Equação 3.2.

Quando há variáveis explicativas categóricas no estudo, tem-se interesse em comparar os efeitos das categorias no tempo, estimando-se as curvas de $S(t)$ via Kaplan-Meier para cada grupo. Entretanto, deve-se utilizar testes de hipóteses para verificar se as diferenças encontradas são estatisticamente significativas (COLOSIMO; GIOLO, 2006). Dentre os testes não paramétricos, os mais utilizados são os testes *logrank* (MANTEL, 1966) e a generalização do teste Wilcoxon proposto por (GEHAN, 1965).

3.1 Teste *logrank*

O teste *logrank* é um dos mais utilizados para comparação de curvas de sobrevivência. Diferencia-se do teste de Wilcoxon, que utiliza peso igual ao número de indivíduos sob risco, consequentemente, apresenta maior penalização no início do acompanhamento, ao invés de colocar o mesmo peso para qualquer que seja o tempo de seguimento como no *logrank*. Além disso, o teste é indicado quando a razão das funções risco dos grupos é aproximadamente constante (COLOSIMO; GIOLO, 2006).

A estatística do teste, obtida de forma análoga ao teste de Mantel-Haenzel, é definida como a diferença entre o número observado de falhas em cada grupo e uma quantidade, podendo ser o número esperado de falhas sob a hipótese nula. Assim, considere as funções de sobrevivência $S_1(t), ..., S_r(t)$, e $t_i, i = 1, ..., k$ os tempos de falhas distintos da amostra formada pela combinação das r amostras e suponha que no tempo t_j ocorram d_j falhas e que n_j indivíduos estejam sob risco em um tempo imediatamente inferior a t_j na amostra combinada e, respectivamente, d_{ij} e n_{ij} na amostra $i; i = 1, .., r$ e $j = 1, ..., k$. Além disso, considere a tabela de contingência (Tabela

Tabela 2 – Tabela de contingência gerada no tempo t_j

	Grupos			Totais
	1	...	r	
Falha	d_{1j}	...	d_{rj}	d_j
Não Falha	$n_{1j} - d_{1j}$	...	$n_{rj} - d_{rj}$	$n_j - d_j$
Totais	n_{1j}	...	n_{rj}	n_j

Nota: Adaptada de Colosimo e Giolo (2006), página 56.

2) que apresenta os dados para cada tempo de falha t_j (MANTEL, 1966; COLOSIMO; GIOLO, 2006).

A distribuição conjunta de $d_{2j}, ..., d_{rj}$ é uma hipergeométrica multivariada, dada a ocorrência de falha e censura até o tempo t_j e ao número de falhas no tempo t_j, é expressa por:

$$\frac{\prod_{i=1}^{r} \binom{n_{ij}}{d_{ij}}}{\binom{n_j}{d_j}}$$

Então, a média de d_{ij} é $w_{ij} = n_{ij} d_j n_j^{-1}$ e as, respectivas, variâncias e covariâncias são:

$$(V_j)_{ii} = n_{ij}(n_j - n_{ij}) d_j (n_j - d_j) n_j^{-2} (n_j - 1)^{-1}$$

e

$$(V_j)_{il} = -n_{ij} n_{lj} d_j (n_j - d_j) n_j^{-2} (n_j - 1)^{-1}.$$

Logo, tem-se a estatística:

$$v = \sum_{j}^{k} v_j,$$

em que, $v_j' = (d_{2j} - w_{2j}, ..., d_{rj} - w_{rj})$, com média igual a zero e matriz de variâncias-covariâncias V_j, com $(V_j)_{ii}, i = 1, ..., r$ e $(V_j)_{il}, i, l = 2, ..., r$.

Supondo que as k tabelas de contingência são independentes, então, $V = V_1 + ... + V_k$. Assim, o teste aproximado para a igualdade das r funções de sobrevivência pode ser pela estatística:

$$T = v' V^{-1} v,$$

sob H_0 : igualdade das curvas, segue a distribuição qui-quadrado com $r - 1$ graus de liberdade (χ^2_{r-1}).

Além disso, os efeitos das categorias no tempo podem ser estimados por modelos paramétricos, em que é associado ao T (tempo de falha) uma distribuição probabilística (COLLETT, 2003). Outro modelo, bastante popular na área médica é o modelo semiparamétrico de Cox, em que, a proporcionalidade dos riscos é a suposição essencial para a sua utilização.

3.2 Modelo de Cox

O modelo semiparamétrico de Cox, também denominado modelo de riscos proporcionais, é um dos mais populares dentre as metodologias estatísticas na área da saúde, quando se tem como objetivo o estudo do tempo até a ocorrência de um evento de interesse. Para Kleinbaum e Klein (2012), a preferência pela metodologia deve-se a robustez do modelo, isto é, os resultados obtidos são muito próximos aos obtidos via modelos paramétricos adequados. Enquanto que, para Colosimo e Giolo (2006), deve-se a presença do componente não paramétrico na formulação, que torna o modelo mais flexível. Além disso, a interpretação dos resultados é mais intuitiva, pois se dá em termos da razão dos riscos estimados.

A formulação do modelo de regressão de Cox é dada por:

$$\lambda(t) = \lambda_0(t) \exp\{\mathbf{X}'\boldsymbol{\beta}\} = \lambda_0(t) \exp\{\beta_1 x_1 + ... + \beta_p x_p\}$$

em que, $\mathbf{X} = (x_1, ..., x_p)$ é o vetor de covariáveis e λ_0, componente não-paramétrico, é função não negativa do tempo t, não especificada e denominada de função de risco basal. Enquanto que, $\exp\{\mathbf{X}'\boldsymbol{\beta}\}$ é o componente paramétrico, sendo o principal interesse de estimação, em especial, o vetor de parâmetros $\boldsymbol{\beta}$.

A suposição para uso do modelo é caracterizada pela razão de taxa de falha entre dois indivíduos distintos ser constante no tempo, isto é,

$$\frac{\lambda_i(t)}{\lambda_j(t)} = \frac{\lambda_0(t) \exp\{\mathbf{X}_i'\boldsymbol{\beta}\}}{\lambda_0(t) \exp\left\{\mathbf{X}_j'\boldsymbol{\beta}\right\}} = \exp\left\{\mathbf{X}_i'\boldsymbol{\beta} - \mathbf{X}_j'\boldsymbol{\beta}\right\},$$

tal que, o resultado não depende do tempo t.

A estimativa dos coeficientes β's, efeito das covariáveis em relação à taxa de falha, é obtida a partir do método da máxima verossimilhança parcial proposta por Cox (1975), pois pela presença do componente não-paramétrico (λ_0) a metodologia usual de estimação não é adequada. A ideia é obter os valores ($\hat{\beta}$'s) que maximizam a função de verossimilhança $L(\boldsymbol{\beta})$.

O método máxima verossimilhança parcial, a fim de eliminar λ_0 da equação, considera a probabilidade condicional da i-ésima observação vir a falhar no tempo t_i, sabendo que as observações estão sob risco em t_i. E também, considera-se que em uma amostra de tamanho n, existam $k \leq n$ falhas distintas nos tempos $t_1 < t_2 < ... < t_k$. Logo, a função de verossimilhança parcial é expressa por:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{k} \frac{\exp\{\mathbf{X}_i'\boldsymbol{\beta}\}}{\sum_{j \in R(t_i)} \exp\left\{\mathbf{X}_j'\boldsymbol{\beta}\right\}} = \prod_{i=1}^{n} \left(\frac{\exp\{\mathbf{X}_i'\boldsymbol{\beta}\}}{\sum_{j \in R(t_i)} \exp\left\{\mathbf{X}_j'\boldsymbol{\beta}\right\}} \right)^{\delta_i},$$

em que, $R(t_i)$ é o conjunto dos índices das observações sob risco no tempo t_i e δ_i é o indicador de falha (LAWLESS, 2011). Como esta função assume que os tempos são contínuos, então espera-se que não haja empates. Dessa forma, nestas ocasiões utiliza-se a aproximação proposta

por Breslow (1972), expressa por:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{k} \frac{\exp\{\mathbf{s}_i'\boldsymbol{\beta}\}}{\left[\sum_{j\in R(t_i)} \exp\left\{\mathbf{X}_j'\boldsymbol{\beta}\right\}\right]^{d_i}},$$

em que, $\mathbf{s}_i'$ é o vetor formado pela soma das respectivas p covariáveis das observações que ocorreram empate, ou seja, mesmo tempo de falha no tempo t_i ($i = 1, ...k$), e d_i é o número de falhas no tempo t_i. Em contrapartida, se a presença de empates for grande, então, é mais adequado utilizar o modelo de Cox para dados agrupados.

Os efeitos das p covariáveis estimados $\hat{\boldsymbol{\beta}}$ determinam o comportamento da taxa de risco, sendo assim, a interpretação é dada por: $\exp\{\beta_l\}, l = 1, ..p$. Desta forma, o risco de morte (ou evento de interesse) de um indivíduo dada uma característica é $\exp\{\beta_l\}$ vezes maior do que indivíduo com uma característica de referência (padrão). Considerando uma covariável quantitativa, o aumento ou diminuição do risco é dado pelo aumento em uma unidade desta característica (COLOSIMO; GIOLO, 2006).

Para verificar a adequabilidade do modelo, tem-se que testar a suposição básica de taxas de falha proporcionais, a partir, principalmente, dos resíduos Schoenfeld, podendo ser graficamente ou por meio de um teste estatístico. Graficamente, avalia-se se existe tendência acentuada da curva ao longo do tempo, e pelo teste, tem-se interesse em não rejeitar a hipótese nula (H_0): correlação igual a zero entre os resíduos e o tempo.

3.2.1 Avaliação da proporcionalidade dos riscos

Método gráfico - risco acumulado

A avaliação é descritiva por meio de um gráfico. Para isso, deve-se dividir os dados em estratos, de acordo com alguma variável e estimar $\hat{\Lambda}_{0j}$, risco acumulado, para cada estrato. A suposição de proporcionalidade dos riscos é satisfeita se as curvas do $log(\hat{\Lambda}_{0j})$ versus t não se cruzarem, ou seja, manterem distância aproximadamente constantes no tempo. Para cada variável que se deseja avaliar, deve-se fazer um novo processo de estratificação e estimação. Nas situações em que as variáveis são contínuas, uma boa estratégia é agrupar os valores em classes (COLOSIMO; GIOLO, 2006).

Método gráfico - Resíduos de Schoenfeld

Considere que se o i-ésimo indivíduo com o vetor de covariáveis $x_i = (x_{1i}, ..., x_{pi})$ falhar, então, para cada indivíduo tem-se um vetor de resíduos de Schoenfeld $r_i = (r_{1i}, ..., r_{pi})$ definido por:

$$r_{iq} = x_{iq} - \frac{\sum_{j\in R(t_i)} x_{jq} exp\{x_j'\hat{\boldsymbol{\beta}}\}}{\sum_{j\in R(t_i)} exp\{x_j'\hat{\boldsymbol{\beta}}\}},$$

sendo d, o número de indivíduos que falharam, $R(t_i)$ é o conjunto dos índices das observações sob risco no tempo t_i e $\boldsymbol{\beta}$, o vetor de parâmetros (COLOSIMO; GIOLO, 2006).

Usualmente, a padronização dos resíduos de Schoenfeld é utilizada para avaliar a suposição de riscos proporcionais e é expressa por:

$$\mathbf{s}_i^* = [I(\hat{\boldsymbol{\beta}})]^{-1} \times \mathbf{r}_i,$$

em que, $I(\hat{\boldsymbol{\beta}})$ é a matriz de informação observada (THERNEAU; GRAMBSCH, 2000; COLOSIMO; GIOLO, 2006). E ainda, Therneau e Grambsch (2000) consideraram o modelo:

$$\lambda(t) = \lambda_0(t) \exp\left\{\mathbf{X}'\boldsymbol{\beta}(\boldsymbol{t})\right\},$$

com a restrição de proporcionalidade dos riscos $\boldsymbol{\beta}(\boldsymbol{t}) = \boldsymbol{\beta}$. Assim, caso $\boldsymbol{\beta}(\boldsymbol{t})$ não seja constante, o risco das covariáveis pode variar com o tempo. Portanto, a suposição de riscos proporcionais é válida se o gráfico de $\mathbf{s}_i^* + \hat{\boldsymbol{\beta}}_q(t)$ versus t, para $q = 1, ..., p$, tiver inclinação aproximadamente zero.

Teste de hipóteses - Resíduos de Schoenfeld

Considere $g_q(t) = g(t)$, uma função do tempo, e a estatística do teste,

$$T = \frac{(g - \bar{g})' S^* I S^{*'} (g - \bar{g})}{d \sum_k (g_k - \bar{g})^2},$$

em que, I é a matriz de informação observada, d é o número de falhas e $S^* = dRI^{-1}$, R a matriz $d \times p$ dos resíduos de Schoenfeld não padronizados. As hipóteses a serem testadas são: H_0: os riscos são proporcionais versus H_1: os riscos não são proporcionais. T tem distribuição qui-quadrado com p graus de liberdade. Valores de $T > \chi^2_{p;1-\alpha}$ indicam haver evidências contra a suposição de riscos proporcionais.

4 MODELOS DE PREDIÇÃO

Modelos de predição são comumente utiliados para prever o comportamento futuro a partir da análise de dados históricos e atuais. Dessa forma, os dados são coletados, um modelo estatístico é formulado, as previsões são feitas e o modelo é validado. Na área da saúde, modelos preditivos são definidos como a combinação de uma série de características para prever um resultado diagnóstico ou prognóstico. Estes modelos são utilizados para investigar a relação entre resultados futuros ou desconhecidos e estados de saúde, no início ou durante o estudo, entre pessoas com condições específicas (STEYERBERG, 2009; LEE; BANG; KIM, 2016). Também podem fornecer a probabilidade do paciente de ter ou desenvolver uma determinada doença. A metodologia preditiva pode ser obtida por algoritmos de aprendizado de máquina, pois o método tem como foco a predição.

A aprendizado de máquina (AM) é um método do ramo da inteligência artificial que busca padrões, tal que, uma máquina pode aprender com dados, identificar padrões e tomar decisões. Com isso, os algoritmos são capazes de analisar dados complexos e maiores com rapidez e precisão (GOLLAPUDI, 2016).

Existem dois tipos principais: supervisionados e não supervisionados. Modelos de aprendizado supervisionado são aqueles em que um modelo é pontuado e ajustado em relação a algum tipo de quantidade conhecida, isto é, quando tenta prever uma variável dependente a partir de uma lista de variáveis independentes. Enquanto que, modelos de aprendizado não supervisionados são aqueles que observa padrões e informações dos dados enquanto determina o próprio parâmetro de ajuste de quantidade, e, geralmente, a variável de interesse é desconhecida, ou seja, as informações não são rotuladas e não há treinamento dos dados (GOLLAPUDI, 2016; BURGER, 2018).

No processo de aprendizado é levado em conta o tipo de variável a ser estudada, isto é, quando a variável de interesse é quantitativa, tem-se problemas de regressão, quando qualitativa, classificação.

Em geral, métodos estatísticos são utilizados pelo pesquisador focando na inferência, a fim de entender como fatores influenciam no comportamento da população em estudo. Entretanto, os métodos de aprendizado de máquina têm como foco a predição, ou seja, prever a variável de interesse com base nas variáveis explicativas. Breiman (2001) discute melhor essas duas vertentes. A cultura com foco na inferência (*data modeling culture*) tem como prioridade, a busca

por modelos corretos, sendo necessário realizar o teste de qualidade de ajuste e a análise de resíduos. Em contrapartida, na cultura voltada para predição (*algorithmic modeling culture*), o modelo não precisa ser o correto, pois a característica mais imprescindível é a criação de um algoritmo preditivo preciso.

Os principais objetivos da aprendizado de máquina são: construir uma função $g : x \to y$, com a finalidade de obter boas predições, ou seja, dadas m novas observações independentes e indenticamente distribuídas observa-se $g(x_{n+1}) \approx y_{n+1}, ..., g(x_{n+m}) \approx y_{n+m}$; e quantificar a qualidade de predição de g, por meio do risco. Por exemplo, para classificação, o risco pode ser definido como $P(g(x) \neq y)$, e regressão, $E[(g(x) - y)^2]$ (BREIMAN, 2001).

Gollapudi (2016) destaca três fases para realizar aprendizado de máquina:
1) treinamento: a maior parte dos dados são utilizados para obter diferentes modelos a fim de encontrar o de menor risco;
2) validação e teste: esta fase tem por objetivo medir o quão bom são os modelos de aprendizado que foram treinados, definir o melhor modelo e estimar as propriedades dele, como medidas de erro, precisão e outras; e
3) aplicação.
Além disso, realizar o pré-processamento, conjunto de técnicas para transformar dados brutos em formatos eficientes, antes das etapas é crucial e pode influenciar nos resultados.

Treinamento, Validação e Teste

Selecionar um modelo a partir de uma amostra $(x_1, y_1), ...(x_n, y_n)$ acarreta em um super-ajuste, isto é, ao ajuste perfeito, pois o modelo foi escolhido de modo a ajustar bem. Como solução, pode-se dividir aleatoriamente o conjuto de dados em dois: validação e treinamento. Comumente, o conjunto de validação representa 25% a 30%, enquanto que, o conjuto treinamento, 70% a 75% dos dados (IZBICKI; SANTOS, 2020; MUELLER; MASSARON, 2021).

Se muitos métodos de predição forem avaliados no conjunto de validação, isto também pode levar a um ajuste perfeito ao conjunto de validação, isto é, pode ocasionar à escolha de um modelo que tem bom desempenho no conjunto de validação, mas não em novas observações. Desta forma, sugere-se dividir o conjunto de dados em três partes: treinamento (70%), validação (20%) e teste (10%). Assim, o conjunto de treinamento será utilizado para estimar os parâmetros do modelo; o conjunto de validação, para estimar o erro quadrático médio, denominado risco observado; e o conjunto de teste é usado para estimar o erro do melhor estimador da regressão determinado com a utilização do conjunto de validação (IZBICKI; SANTOS, 2020; MUELLER; MASSARON, 2021).

O processo de divisão ou reamostragem dos dados pode ser realizada pelos métodos validação cruzada *holdout*, validação cruzada *k-fold* e *bootstrap*.

Validação cruzada *holdout*

O método se resume em dividir o conjunto de dados em treinamento e validação ou treinamento, validação e teste. A implementação é considerada simples e computacionalmente não custosa.

A divisão dos dados pode dar-se por meio dos métodos de amostragem aleatória simples: mantém a tendência dos dados; aleatória estratificada: garante a correta distribuição dos grupos; e amostragem não aleatória: quando deseja-se que o conjunto de teste seja composto por observações mais recentes. Entretanto, recomenda-se que a divisão dos dados seja realizada aleatoriamente. Para isso, utiliza-se um gerador de números aleatórios para escolher quais amostras constituirão cada grupo (treinamento, validação e teste) (KUHN; JOHNSON, 2019).

Na prática, a validação por *holdout* não é muito utilizada, pois a precisão da avaliação da qualidade do algoritmo pode não ser precisa, já que depende de ter escolhido um bom subconjunto de teste. Em outras situações o tamanho da amostra é insuficiente para divisão do conjunto de dados, gerando modelos mal ajustados e aproximações para o erro do teste superestimadas (FERREIRA, 2018).

Validação cruzada *k-folds*

A validação cruzada por *k-fold* consiste em dividir os dados em k subconjuntos iguais (*folds*). Em seguida, o modelo AM é treinado utilizando $k-1$ subconjuntos, e a parte restante, que não foi utilizada no treinamento, é destinada à validação. Esse processo é repetido k vezes, com um subconjunto diferente reservado para validação. Por fim, os resultados são combinados obtendo a média dos erros obtidos. Diferente da metodologia apresentada anteriormente, esta minimiza os problemas de ajuste e obtenção de erros superestimados (FERREIRA, 2018).

Bootstrap

O método *bootstrap* foi apresentado de forma sistematizada por Efron em 1979. Definido como método de reamostragem que pode ser usado para quantificar a incerteza associada a propriedades de estimadores ou método de aprendizado de máquina. Esta metodologia é aplicável a diversos métodos estatísticos e facilita a obtenção da variância, quando não é tão simples determinar ou não fornecido pelo software de análise estatística (MAYER, 2021).

Suponha $x_1, ..., x_n$ uma amostra aleatória de tamanho n de uma população. Neste caso, espera-se que essa amostra possua características similares a população. Com isso, o método *bootstrap* considera a amostra como "população" e realiza amostragem a partir dela. Logo, coleta-se aleatoriamente, com reposição, B amostras de tamanho n. Para cada uma das $b = 1, ..., B$ amostras, pode-se obter a estimativa de interesse e o respectivo desvio padrão. Portanto, a

estimativa pontual *bootstrap* é o valor médio (FERREIRA, 2018; MAYER, 2021),

$$\bar{\hat{\theta}} = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}^{(b)}.$$

Avaliação do desempenho dos modelos

Uma maneira comum de avaliar a precisão preditiva dos modelos nos dados de sobrevivência é considerar o risco relativo de um evento para diferentes observações em vez dos tempos de sobrevivência absolutos para cada observação. Isso pode ser feito computando a probabilidade de concordância ou o índice de concordância (índice C) proposto por Harrell et al. (1982).

Considere duas observações e seus respectivos valores de previsão, $(y_1, \hat{y}_1)$ e $(y_2, \hat{y}_2)$, em que, y_i e $\hat{y}_i$ representam o tempo real de observação e o valor previsto, respectivamente, e que para o paciente i, nosso modelo de risco atribui uma pontuação de risco η_i. A probabilidade de concordância entre eles pode ser calculada como:

$$C = P(\hat{y}_1 > \hat{y}_2 | y_1 \geq y_2).$$

Mais detalhadamente, considere que para cada par de indivíduos i e j ($i \neq j$), seja observado os escores de risco η e os tempos até o evento de interesse. Então, se ambos T_i e T_j não forem censurados, significa que observou-se para ambos indivíduos o evento (por exemplo, o óbito). Logo, o par (i, j) é um par concordante se $\eta_i > \eta_j$ e $T_i < T_j$, e é um par discordante se $\eta_i > \eta_j$ e $T_i > T_j$. Caso contrário, se os tempos são censurados, então, não se sabe o desfecho dos indivíduos. Com isso, os tempo não serão considerados no cálculo. Nas situações em que apenas um dos tempos é censurado, isto é, observa-se o evento para apenas um indivíduo. Então, se $T_j < T_i$, não considera-se esse par no cálculo e se $T_j > T_i$, considera-se, pois sabe-se qual indivíduo o evento foi observado. Portanto, (i, j) é um par concordante se $\eta_i > \eta_j$, e é um par discordante se $\eta_i < \eta_j$ (SCHMID; WRIGHT; ZIEGLER, 2016; TAY, 2019).

Portanto, o índice C é dado por:

$$C = \frac{\#\text{ pares concordantes}}{\#\text{ pares concordantes} + \#\text{ pares discordantes}}.$$

Assim, valor de $c = 0,5$ corresponde a uma previsão não informativa, enquanto $c = 1$ corresponde à associação perfeita (SCHMID; WRIGHT; ZIEGLER, 2016).

Dentre os métodos que podem ser modelados a partir do algoritmo de aprendizado de máquina, neste trabalho, serão utilizadas as técnicas baseadas em árvores de classificação e regressão, redes neurais artificiais e máquinas de vetores de suporte.

4.1 Árvores de Classificação e Regressão

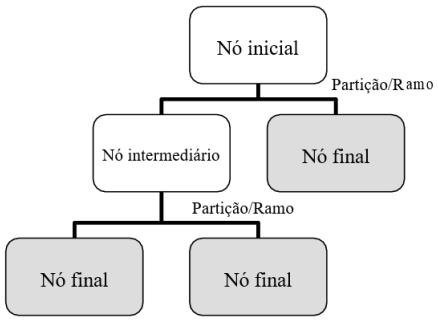
A metodologia de partição recursiva é um método de aprendizado de máquina não paramétrico utilizado como modelo de predição desenvolvido por Morgan e Sonquist (1963), e enquanto que, a termologia e o algoritmo de árvores de classificação e regressão (CART) foram difundidos por Breiman et al. (1984). Este método estabelece a relação entre a variável resposta quantitativa (regressão) ou qualitativa (classificação) em função de um conjunto de variáveis explicativas.

Taconeli (2008) divide o processo de construção de árvore em quatro etapas: execução de um critério de partição das amostras, aplicação do processo de poda, seleção do melhor modelo e classificação dos nós finais. Na filosofia do algoritmo CART prefere-se em escolher a melhor partição possível (GOLLAPUDI, 2016).

O CART se baseia em dividir a amostra estudada em subamostras homogêneas, ou seja, escolhe-se em cada nó a variável que mais discrimina as subamostras resultando na bifurcação da árvore em dois ramos. Para cada ramo criado escolhe-se outro nó baseado na próxima variável mais discriminante, e o processo segue assim por diante (LUCAS, 2011).

Taconeli (2008) apresenta um exemplo que caracteriza bem os componentes de uma árvore, entre eles, o nó inicial, que é a variável mais discriminante dentro da amostra original, os nós intermediários que são as variáveis que melhor discrimina e originam novas subamostras e os nós finais que são as variáveis com menor capacidade de discriminar e reunem subamostras não partidas (Figura 2).

Figura 2 – Ilustração de uma árvore de classificação/regressão



Fonte: Taconeli (2008).

De forma geral, suponha uma matriz que armazena os dados $M \in M_{n,p+1}$ com n observações e p variáveis preditoras $x_1, ..., x_p$ e Y, variável resposta com k classes distintas. O principal objetivo é criar uma partição do espaço de dimensões da variável de previsão para associar um valor de Y a cada partição $m_1, ..., m_l$.

O custo-complexidade de uma árvore A (BREIMAN et al., 1984) é definido por,

$$R_\alpha(A) = \sum_{h \in \bar{A}} R(h) + \alpha|\bar{A}|,$$

e é utilizado para escolher o tamanho da árvore. Para um parâmetro de complexidade não negativo α, em que $R(h)$ é a heterogeneidade dentro do nó h e sua fórmula depende do critério de divisão escolhido, considerando o modelo de risco relativo, e $|\bar{A}|$ é o número de nós terminais de A, ou seja, a complexidade de A.

A regra de divisão utilizada nas árvores de classificação é o índice de Gini, que é uma medida da variância total nas k classes, definida por:

$$G = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk}),$$

em que, p_{mk} representa a proporção de observações de treino na m-ésima partição que são da k-ésima classe.

Para árvores de regressão, Moisen (2008) apresenta duas medidas que podem ser utilizadas como regra de divisão: mínimos quadrados, obtido por meio da média, e mínimos desvios absolutos, obtido por meio da mediana.

O processo de construção da árvore pode produzir boas previsões no conjunto de treinamento, entretanto pela complexidade e particularidade é provável que haja um super-ajuste dos dados, levando a um desempenho preditivo ruim quando aplicado a outros conjuntos de dados. Como estratégia, James et al. (2013) sugerem a construção de uma árvore muito grande, para depois podar e obter uma subárvore que tenha menor taxa de erro de teste. Assim, o tamanho ideal de uma árvore pode ser determinado utilizando um conjunto de testes independentes ou validação cruzada (MOISEN, 2008).

O método CART é uma modelagem não paramétrica que pode ser utilizado como uma alternativa para modelos lineares generalizados, regressão linear múltipla, regressão logística, análise de sobrevivência, análise discriminante, correlação canônica ou análise de agrupamentos, entre outros (TACONELI, 2008).

Além do mais, as árvores podem ser exibidas graficamente e dessa forma serem facilmente interpretadas mesmo por um não especialista e podem lidar com preditores qualitativos sem a necessidade de criar variáveis "*dummy*" (JAMES et al., 2013).

Outra vantagem da metodologia CART é a robustez na incidência de pontos discrepantes, pois o método é flexível, não impondo restrições quanto à natureza e à distribuição de probabilidade das variáveis. Quando se tem um grande número de informações das quais se aprende e extrai conhecimento (aprendizado de máquina), que não são tão fáceis de serem manipuladas por metodologias tradicionais, a utilização de métodos multivariados é indicada, como por exemplo, o método de árvores (TACONELI, 2008).

4.1.1 Árvores de Sobrevivência

A construção de árvores de sobrevivência refere-se ao método de árvores de classificação e regressão (BREIMAN et al., 1984), aplicado à análise de dados de sobrevivência, explorado inicialmente por Gordon e Olshen (1985).

A intuição básica por trás dos modelos da árvore é particionar recursivamente os dados com base em um critério de divisão particular, e os objetos que são semelhantes entre si, que com base no evento de interesse serão colocados no mesmo nó (WANG; LI; REDDY, 2019).

A principal diferença entre uma árvore de sobrevivência e a árvore de decisão padrão está na escolha do critério de divisão, que leva em consideração a censura presente nos dados em estudo. Nos últimos anos, diferentes metodologias foram propostas para dados censurados, entre eles, baseado na distância Wasserstein entre as curvas de Kaplan-Meier (GORDON; OLSHEN, 1985), na estatística *logrank* (CIAMPI et al., 1986), nos resíduos de *martingale* (THERNEAU; GRAMBSCH; FLEMING, 1990), no tempo médio restrito de sobrevivência (JIN et al., 2004) e no custo da divisão da variável a partir da estatística *logrank* (JIN; LU, 2011).

Dentre os métodos disponíveis, tem-se as árvores de risco relativo (*relative risk trees*) proposto por LeBlanc e Crowley (1992), que incorpora o modelo de riscos proporcionais, em que a função de risco no tempo t, para um indivíduo com vetor de covariável $\mathbf{X}$, é dada por:

$$\lambda(t|\mathbf{X}) = \lambda_0(t)g(\mathbf{X})$$

tal que, $g(\mathbf{X}) \geq 0$ e $\lambda_0(t)$ é a função risco basal. Normalmente, $g(\mathbf{X}) \geq 0$ é uma função log-linear de um vetor de parâmetros. Entretanto, para a metodologia a ser apresentada, o objetivo é obter estrutura de árvores que representam a função de risco relativo, $g(\mathbf{X}) \geq 0$.

Em resumo, a estimativa da verossimilhança completa é obtida iterativamente, com a finalidade de obter a árvore que minimiza o erro de predição estimado. Contudo, a construção da árvore utiliza validação cruzada para estimar esse erro.

As inferências para o modelo de riscos proporcionais são baseadas na verossimilhança parcial. A verossimilhança completa da amostra de aprendizado para a árvore A pode ser expressa como (LEBLANC; CROWLEY, 1992):

$$L = \prod_{h \in \tilde{A}} \prod_{i \in O_h} \left(\lambda_0(t_i)\theta_h\right)^{\delta_i} \exp\{-\Lambda_0(t_i)\theta_h\},$$

em que, $\tilde{A}$ é o conjunto de nós terminais; O_h é o conjunto das observações que estão no nó h; (t_i, δ_i) é o vetor do tempo observado e indicador de falha para o indivíduo i; $\lambda_h(t_i)$ e $\Lambda_h(t_i)$ são as funções risco e risco acumulado para o nó h, respectivamente; e $\Lambda_0(t)$ é o risco basal acumulado. Sendo assim, dado o risco basal acumulado, o estimador de máxima verossimilhança de $\{\theta_h : h \in \tilde{A}\}$ é expresso por:

$$\hat{\theta}_h = \frac{\sum_{i \in O_h} \delta_i}{\sum_{i \in O_h} \Lambda_0(t_i)}.$$

Para obter o estimador de Breslow do risco basal acumulado dada a estimativa $\{\theta_h : h \in \tilde{A}\}$, um procedimento alternativo de estimação pode ser usado para estimar Λ_0 e $\{\theta_h : h \in \tilde{A}\}$. Assim, num primeiro momento, para a iteração j,

$$\hat{\Lambda}_0^j = \sum_{i:t_i \leq t} \frac{\delta_i}{\sum_{h \in \tilde{A}} \sum_{i:t_i \geq t, i \in O_h} \hat{\theta}_h^j}, \tag{4.1}$$

é calculado usando as estimativas atuais, $\hat{\theta}_h^j$ de θ_h. Em seguida, a estimativa $\hat{\theta}_h^{j+1}$ de θ_h,

$$\hat{\theta}_h^{j+1} = \frac{\sum_{i \in O_h} \delta_i}{\sum_{i \in O_h} \hat{\Lambda}_0^j(t_i)}, \tag{4.2}$$

é calculada usando a estimativa atual $\hat{\Lambda}_0^j(t)$. Os passos são repetidos até a convergência. Importante ressaltar que, somente a primeira iteração é utilizada para partição recursiva para construir e podar (selecionar o tamanho) da árvore.

A árvore selecionada, denominada sub-árvore otimamente podada A_α, é aquela que minimiza a medida de custo-complexidade α definida por Breiman et al. (1984), como na metodologia CART. Isto é, matematicamente, A_α é escolhida dentre todas as sub-árvores se $R_\alpha(A_\alpha) = min\{A' \leq A\} R_\alpha(A')$.

Existem situações em que todas as observações no nó h são censuradas na sub-amostra usada para construir a árvore, neste caso, $\hat{\theta}_h = 0$. Além disso, a estimativa de validação cruzada da deviance esperada é infinita, quando os nós apresentam somente censura. Logo, sugere-se substituir nós com nenhuma falha observada por 0,5. Então, o estimador de θ_h para validação cruzada é dado por:

$$\hat{\theta}_h = \frac{1}{2 \sum_{i \in O_h} \hat{\Lambda}_0^1(t_i)},$$

para um nó h que não têm falhas. Depois de escolher uma árvore $A(\alpha^*)$ minimizando a estimativa da deviance esperada pós validação cruzada, estimadores de máxima verossimilhança do risco relativo entre os nós são obtidos pelo processo iterativo apresentado (Equações 4.1 e 4.2) (LEBLANC; CROWLEY, 1992).

4.2 Redes Neurais Artificiais

As redes neurais artificiais (RNA) são modelos estatísticos não lineares inspirados pelo comportamento biológico de neurônios cerebrais e suas interconexões. Desta forma, as RNA são construídas como um denso conjunto de unidades conectadas entre si (FRIEDMAN et al., 2001; GERRISH, 2018) e representadas graficamente pela Figura 3.

No diagrama de rede neural (Figura 3), os neurônios são representados pelos nós (círculos) e os pesos β de suas respectivas conexões, pelas setas. Sendo que, os nós à esquerda representam as covariáveis em estudo (entrada da rede), os nós centrais pertencem a camada oculta e representam a transformação dos nós entrada da rede (GERRISH, 2018). Neste caso, considere $\mathbf{X} = (x_1, x_2, x_3)$,

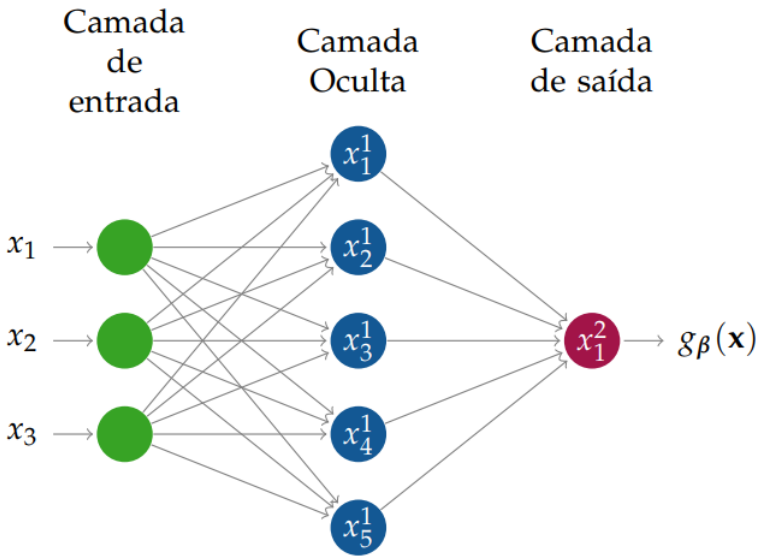
vetor de covariáveis de entrada, então um dado neurônio j da camada oculta é calculado como (IZBICKI; SANTOS, 2020; HAYKIN, 2009):

$$x_j^1 := f\left(\beta_{0,j}^0 + \sum_{i=1}^{3} \beta_{i,j}^0 x_i^0\right),$$

em que, $x_i^0 = x_i$ para $i = 1, 2, 3$, tal que, f é denominada função de ativação [1], determinada pelo pesquisador. Obtido x_j^1 para cada neurônio j da camada oculta, onde o índice sobrescrito representa a camada. Para este exemplo, o neurônio da camada de saída do modelo é dado por:

$$x_1^2 := f\left(\beta_{0,j}^1 + \sum_{i=1}^{5} \beta_{i,1}^1 x_i^1\right) =: g_\beta(\mathbf{X}).$$

Figura 3 – Diagrama de rede neural com uma rede oculta



Fonte: Izbicki e Santos (2020).

[1] Funções comumente utilizadas são: identidade, logística, tangente hiperbólica, ReLu (rectified linear) e Leaky ReLun.

Generalizando, considere uma rede com H camadas ocultas, cada uma com d_l neurônios ($l = 1, ..., H$), $\beta_{i,j}^l$ o peso atribuído à conexão entre a entrada i na camada l e a saída j na camada $l+1$, $l = 0, ..., H$, tal que, $l = 0$ denota a camada de entrada, e $H+1$ a camada de saída. Então, a função de regressão para obter o estimador de uma rede neural artificial é expressa por (ROJAS, 2013; IZBICKI; SANTOS, 2020):

$$g(\mathbf{X}) = x_l^{H+1} = f\left(\beta_{0,l}^H + \sum_{i=1}^{d_H} \beta_{i,l}^H x_i^H\right)$$

em que, para todo $l = 1, ..., H$ e $j = 1, ..., d_{l+1}$,

$$x_j^{l+1} = f\left(\beta_{0,j}^l + \sum_{i=1}^{d_l} \beta_{i,j}^l x_i^l\right)$$

Segundo Haykin (2009), pode-se destacar diferentes arquiteturas de redes: redes *feed-forward* de camada única - definido por uma camada de entrada de nós de origem que se projeta diretamente em uma camada de saída de neurônios (nós); redes multicamadas *feedforward* - definido pela presença de uma ou mais camadas ocultas (não é vista diretamente da entrada ou da saída da rede), cujos nós são chamados de neurônios ocultos; e redes recorrentes - distingue-se de uma rede neural *feedforward* por ter pelo menos um loop de *feedback*, sendo comumente utilizado o auto-feedback (retroalimentação), podendo ter ou não camada oculta. O auto-feedback refere-se a uma situação em que a saída de um neurônio é realimentada em sua própria entrada.

Para estimar os coeficientes $\beta_{i,j}^l$, normalmente, utiliza-se o algoritmo *backpropagation*, por meio do gradiente de decrescimento máximo para minimizar o erro quadrático entre os valores de saída da rede neural e os respectivos valores objetivos. Primeiro atualizam-se os coeficientes da última camada, depois atualizam-se os coeficientes da camada anterior a essa e assim por diante. Entretanto, o vetor β que minimiza:

$$EQM(g_\beta) = \frac{1}{n}\sum_{k=1}^{n} \left(g_\beta(\mathbf{X}_k) - y_k\right)^2,$$

não possui solução analítica, por isso, métodos numéricos são utilizados para estimá-lo (FRIEDMAN et al., 2001; HAYKIN, 2009).

Assim, o *backpropagation* consiste em fazer, para $l = H, ..., 0$, a seguinte atualização (ROJAS, 2013):

$$\beta_{i,j}^l \leftarrow \beta_{i,j}^l - \eta \frac{\partial EQM(g_\beta)}{\partial \beta_{i,j}^l}, j = 1, ..., d_{l+1} \; e \; i = 0, ..., d_l,$$

η é a taxa de aprendizado, ou seja, um parâmetro de proporcionalidade que define o comprimento da etapa de cada iteração na direção do gradiente.

4.2.1 Redes Neurais de Sobrevivência

Considere a função de risco do modelo de Cox:

$$h(t, \mathbf{X}) = h_0(t) \exp(\mathbf{X}'\beta),$$

tal que, o vetor de parâmetros β é estimado maximizando a seguinte a função de verossimilhança parcial:

$$L_c(\beta) = \prod_{i \in u} \frac{\exp(\mathbf{X}_i'\beta)}{\sum_{j \in R_i} \mathbf{X}_j'\beta},$$

via método Newton-Raphson.

Substituindo o preditor linear $\mathbf{X}'\beta$ pela saída da rede $g_\beta(\mathbf{X})$, o modelo de riscos proporcionais é dado por $\lambda(t) = \lambda_0(t) \exp[g_\beta(\mathbf{X})]$ e a função a ser maximizada é dada por (FARAGGI; SIMON, 1995):

$$L(\beta) = \prod_{k \in u} \frac{\exp(g_\beta(\mathbf{X}_k))}{\sum_{j \in R_k} g_\beta(\mathbf{X}_j)} = \prod_{k \in u} \frac{\exp\left(f\left(\beta_{0,l}^H + \sum_{i=1}^{d_H} \beta_{i,l}^H x_{i,k}^H\right)\right)}{\sum_{j \in R_k} f\left(\beta_{0,l}^H + \sum_{i=1}^{d_H} \beta_{i,l}^H x_{i,k}^H\right)}.$$

Outros modelos de dados censurados podem ser utilizados na topologia de redes neurais, para isso, requer a substituição de $\mathbf{X}'\beta$ por uma função não linear de rede neural $g_\beta(\mathbf{X})$ e prosseguir com os algoritmos usuais para estimar os parâmetros, como os modelos de tempo de falha acelerado e o modelo de Buckley-James (KATZMAN et al., 2018a).

4.3 Máquinas de Vetores de Suporte

O algoritmo de máquinas de vetores de suporte (MVS) foi desenvolvido por Vapnik e Lerner (1963) como um método de aprendizado para problemas de reconhecimento de padrões. A metodologia foi generalizada obtendo funções não lineares que criam hiperplanos separadores para variáveis de interesse dicotômicas (CORTES; VAPNIK, 1995). Além disso, o algoritmo foi adaptado para regressão com a finalidade de estimar valores reais ($\mathbb{R}$) (VAPNIK, 1999).

A fim de facilitar o entedimento do algoritmo MVS, primeiro será apresentado a metodologia de classificação. Considere n observações compostas pelos pares $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), ..., (\mathbf{x}_n, y_n)$, com $\mathbf{x}_i \in \mathbb{R}^p$ e $y_i \in \{-1, 1\}$, e a função linear dada por:

$$f(\mathbf{X}) := \beta_0 + \beta_1 x_1 + ... + \beta_p x_p.$$

Então, o classificador $g(\mathbf{X})$ pode ser expresso por (FRIEDMAN et al., 2001; IZBICKI; SANTOS, 2020):

$$\begin{cases} \text{Se } f(\mathbf{X}) < 0, g(\mathbf{X}) = -1 \\ \text{Se } f(\mathbf{X}) \geq 0, g(\mathbf{X}) = 1 \end{cases}$$

A função $f(\mathbf{X})$, definida como hiperplano, supõe que as observações sejam linearmente separáveis, dividido-as perfeitamente em duas classes. Desta forma, existe $f(\mathbf{X})$ linear tal que

$f(\mathbf{x}_i) < 0$ se, e somente se, $y_i = -1$. Logo, para todo $i = 1, ..., n$,

$$y_i(\beta_0 + \beta_1 x_{1,i} + ... + \beta_p x_{p,i}) = y_i f(\mathbf{x}_i) > 0.$$

A solução é definida como a linha de separação com maior margem M, ou seja, que determina o maior espaço vazio (distância) entre o limite das classes e os pontos observados. Os pontos utilizados para definir as margens são chamados de vetores de suporte (IZBICKI; SANTOS, 2020; MUELLER; MASSARON, 2021).

Contudo, ao procurar a maior margem M de separação, é levado em consideração que o algoritmo deriva de uma função de classificação de uma amostra de observações, mas que não pode confiar nestas observações. Consequentemente, o algoritmo não é influenciado por pontos discrepantes (ao contrário de uma regressão linear). O algoritmo determina a margem usando apenas os pontos nos limites (MUELLER; MASSARON, 2021).

Se as duas classes não forem linearmente separáveis, pode-se permitir que alguns pontos fiquem do lado errado da margem, acarretando em erros de classificação, com a introdução de variáveis de folga $\xi_i = |y_i - f(x_i)| \geq 0, i = 1, ..., n$. Entretando, os erros de classificação são penalizados. Portanto, $\xi_i = 0$ se i for classificado corretamente, $\xi_i < 1$ se estiver dentro da margem e do lado correto do hiperplano, $\xi_i > 1$ se o ponto de dados estiver do lado errado do hiperplano, e $\xi_i = 1$ se a observação i estiver exatamente no hiperplano (FOUODO et al., 2018).

As definições descritas podem ser ilustradas na Figura 4, considerando um conjunto bidimensional. Nesta representação, os dados são agrupados em duas classes, tal que, os pontos preenchidos são os vetores de suporte e os em vermelho (*misclassified*), foram classificados incorretamente. Finalmente, a linha tracejada representa o hiperplano.

Em termos de um problema de otimização, tem-se como objetivo:

$$\max_{\beta, \beta_0, ||\beta||=1} M$$

sujeito a,

$$y_i(\mathbf{X}'\beta + \beta_0) \geq M(1 - \xi_i), i = 1, ..., n.$$

O problema pode ser reescrito como,

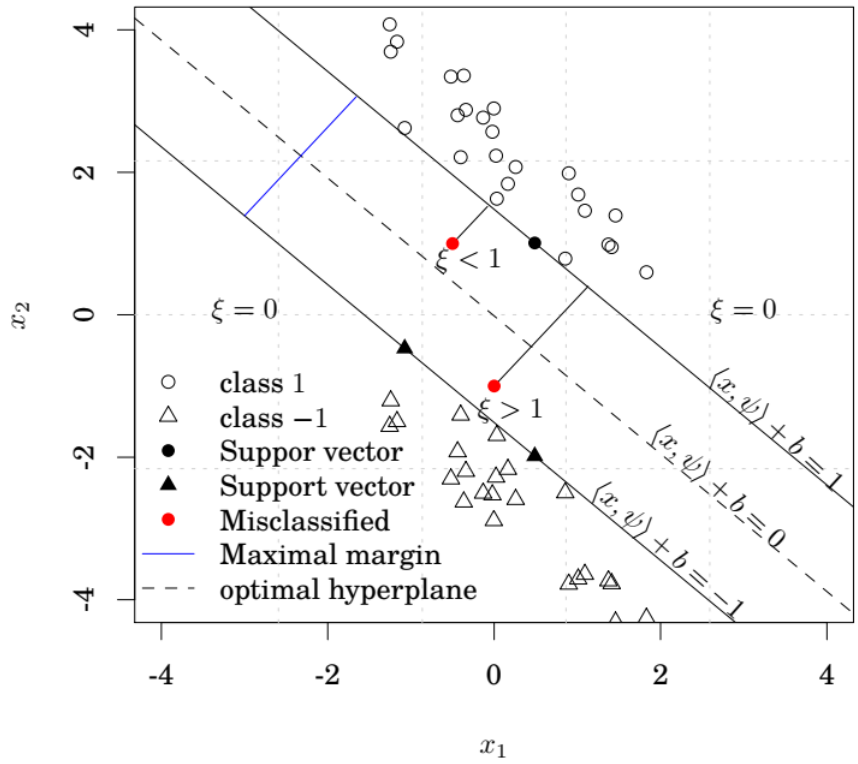
$$\min_{\beta, \beta_0} ||\beta||$$

sujeito a,

$$y_i(\mathbf{X}'\beta + \beta_0) \geq 1 - \xi_i, i = 1, ..., n,$$

$$\xi_i \geq 0, \sum \xi_i \leq constante.$$

Figura 4 – Exemplo de aplicação do MVS destacando os vetores de suporte, a margem e os erros de classificação



Fonte: Fouodo et al. (2018).

Para facilitar a solução, computacionalmente, é convenientemente reescrever o problema como,

$$\min_{\beta,\beta_0,\xi} \frac{1}{2}||\beta||^2 + \gamma \sum_{i=1}^{n} \xi_i$$

sujeito a,

$$-(y_i(\langle \mathbf{x}_i, \beta \rangle + \beta_0) + \xi - 1) \leq 0,$$
$$\xi_i \geq 0, i = 1, ..., n,$$

em que, y e as variáveis de folga ξ_i são desconhecidas, n é o número de observações e $\gamma > 0$ é um parâmetro de regularização ("custo") que controla a margem máxima e as penalidades de classificação incorreta. Desta forma, a solução será obtida transformando o problema de otimização comum em um problema de otimização dual e utilizando multiplicadores de *Lagrange*. Para maiores detalhes, ver seção 12.2.1 do Friedman et al. (2001).

4.3.1 Máquinas de Vetores de Suporte de Sobrevivência

Três principais abordagens foram propostas para resolver problemas de sobrevivência usando MVS: a regressão, a classificação e a híbrida. O foco nesta seção é apresentar a abordagem de regressão apresentada por Shivaswamy, Chu e Jansche (2007). A metodologia baseia-se na ideia de regressão do vetor de suporte (RVS) e tem como objetivo estimar os tempos de sobrevivência observados como valores de resultados contínuos y_i utilizando covariáveis $\mathbf{x}_i$.

Como descrito por Fouodo et al. (2018), considerando as observações censuradas, o tempo até o evento de interesse após a censura é desconhecido e, portanto, previsões maiores que o tempo de censura não precisam ser penalizadas. No entanto, todas as previsões de sobrevivência abaixo do tempo de censura são penalizadas como de costume. Para dados não censurados, os tempos de sobrevivência exatos são conhecidos e, como no RVS padrão, todas as previsões de sobrevivência menores ou maiores do que o tempo de sobrevivência observado são penalizadas.

Assim, a formulação deste problema pode ser expressa por (SHIVASWAMY; CHU; JANSCHE, 2007; FOUODO et al., 2018):

$$\min_{\beta,\beta_0,\xi,\xi^*} \frac{1}{2}||\beta||^2 + \gamma \sum_{i=1}^{n}(\xi_i + \xi_i^*)$$

sujeito a,

$$y_i - \langle \beta, F(\mathbf{x}_i)\rangle - \beta_0 \leq \xi_i,$$

$$\delta_i(\langle \beta, F(\mathbf{x}_i)\rangle + \beta_0 - y_i) \leq \xi_i^*$$

$$\xi_i, \xi_i^* \geq 0, i = 1, ..., n,$$

em que, δ é o indicador de falha e F é uma função que mapeia covariáveis observadas para o espaço de recursos. No caso de nenhuma censura, ou seja, $\delta_i = 1$, as restrições de desigualdade são as mesmas do problema RVS clássico. No entanto, se ocorrer censura, a segunda restrição é reduzida para $\xi_i^* \geq 0$.

O espaço de recursos, em comparação com o espaço original, no qual os dados são observados, é um espaço de dimensão superior no qual os pontos de dados são projetados para serem separados mais facilmente por um hiperplano. Basta ser capaz de calcular os produtos internos entre os vetores de suporte e os vetores do espaço de recursos, isto é, uma vez que o espaço de recursos implica um espaço dimensional mais alto, o produto interno $\langle \beta, F(\mathbf{x}_i)\rangle$ é calculado usando uma função *kernel* para reduzir o tempo de execução (VAPNIK, 1999; FOUODO et al., 2018).

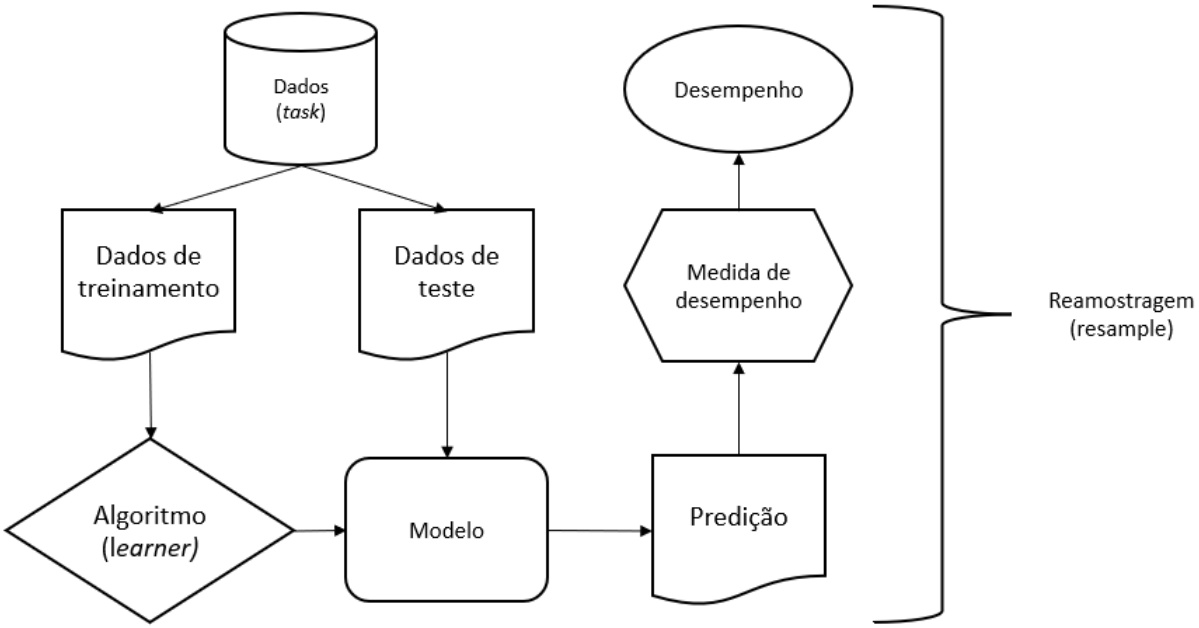
4.4 Aprendizado de Máquina no *software* R

O *software* utilizado para implementar e analisar os dados ao longo desta tese será o *software* estatístico R (R Core Team, 2021), *software* livre para realizar análises estatísticas e construção de gráficos.

Os pacotes principais utilizados são específicos de aprendizado de máquina *mlr* (BISCHL et al., 2016) e *mlr3* (LANG et al., 2019). Este último será o foco desta seção por ser uma extensão mais completa do *mlr*. Além disso, será considerado apenas dados de sobrevivência, para os demais tipos de resposta (variável de interesse), qualitativa (classificação) e quantitativa (regressão), é possível encontrar com facilidade em diversas aplicações disponíveis em trabalhos acadêmicos e sites de busca.

Na Figura 5, apresenta-se o fluxograma dos procedimentos para aplicar a metodologia de aprendizado de máquina e que serão descritos nas próximas seções.

Figura 5 – Fluxograma de um processo de aprendizado de máquina



Fonte: Adaptado de Becker et al. (2021).

4.4.1 *Task*

Task são objetos que contêm os dados e a variável resposta a fim de definir um modelo de aprendizado de máquina. Este objeto é parte da execução do algoritmo de otimização e irá compor uma função como parâmetro. A variável de interesse é o tempo até a ocorrência de um evento ou censura, então, estes serão os parâmetros a serem definidos na função *TaskSurv* de pacote *mlr3proba* (SONABEND et al., 2021).

```
    TaskSurv$new(
    id, backend, time = "time", event = "event", time2,
    type = c("right", "left", "interval", "counting", "interval2", "mstate"
        ))
```

Código R 4.1 – Definição do objeto *task*

Sendo, o id é nome do novo método *task* (*new*), *backend* indica o conjunto de dados; *time*, variável tempo de acompanhamento (censura à direita) ou valor inicial do intervalo de tempo censurado; *event*, variável indicadora de falha; *time2*, valor final do intervalo de tempo censurado (se necessário); e *type*, tipo de censura (SONABEND et al., 2021).

4.4.2 Learner

O objeto *learner* consiste em designar métodos de aprendizado de máquina para treinar e prever modelos para um *task*, podendo definir os valores para os hiperparâmetros diferentes dos predefinidos (padrões).

```
    learner = lrn("nome.learner") #pode acrescentar valores dos
        hiperparametros
    # exemplo
    learner = lrn("surv.rpart")
    learner = lrn("surv.rpart", cp = 0.001)
```

Código R 4.2 – Definição do objeto *learner*

O pacote *mlr3extralearners* dispõe de uma lista de *learners* do *mlr3* e utilizando o comando *list_mlr3learners()* é possível visualizar.

4.4.3 Treino, previsão e avaliação do desempenho

O processo de treino, previsão e avaliação pode ser automatizado com o processo de reamostragem, aqui será apresentado como pode-se dividir o conjunto de dados em subconjuntos de treinamento e teste, composto por 70% e 30% das observações, respectivamente (BECKER et al., 2021).

```
    set.seed(1234)
    train_set = sample(task$row_ids, 0.7 * task$nrow)
    test_set = setdiff(task$row_ids, train_set)

    # treinamento
    learner$train(task, row_ids = train_set)

    # previsao
    prediction = learner$predict(task, row_ids = test_set)

```

```
11    #avaliacao
12    measure = msr("surv.cindex")
13    prediction$score(measure)
```

Código R 4.3 – *Script* - treino, previsão e avaliação do desempenho

4.4.4 Reamostragem

A reamostragem é uma metodologia para criar divisões de treinamento e teste, para auxiliar na avaliação de desempenho do modelo treinado. Dentre as metodologias disponíveis no *mlr3*, tem-se: validação cruzada (e variações), bootstrap e divisão simples (*holdout*) (BECKER et al., 2021).

```
1    # definicao da reamostragem - validacao cruzada \textit{k-folds}
2    resampling = rsmp("cv", folds = 5)
3    rr = resample(task, learner, resampling, store_models = TRUE)
4    # desempenho medio das iteracoes - \textit{cindex}
5    rr$aggregate(msr("surv.cindex"))
6    # grafico dos resultados da amostragem
7    library(mlr3viz)
8    autoplot(rr, measure = msr("surv.cindex"))
```

Código R 4.4 – Avaliação do desempenho via reamostragem (*resampling*)

4.4.5 Comparação de desempenho de modelos

Comparar o desempenho de diferentes *learners* - modelos treinados - é a forma mais comum de apresentação, na literatura, dos resultados em aprendizado de máquina. Esse processo, geralmente, é denominado de *benchmarking* (BECKER et al., 2021).

```
1    library("mlr3verse")
2    # definicao dos parametros - vetores (conjunto) de valores
3    design = benchmark_grid(tasks, learners, resamplings)
4    bmr = benchmark(design)
5    # desempenho medio - cindex
6    bmr$aggregate() # bmr$aggregate(measures)
```

Código R 4.5 – Comparação em modelos - definição do *benchmarking*

4.4.6 Exemplo - dados *Lung*

Nesta seção será apresentada didaticamente um exemplo para facilitar a reprodução da metodologia de aprendizado de máquina para quaisquer dados de análise de sobrevivência visto a dificuldade de encontrar exemplos práticos na literatura. Esse passo a passo é o mesmo utilizado para análise de dados desse trabalho.

O conjunto de dados de câncer de pulmão (*lung*) utilizado no exemplo está disponível no pacote *survival* e apresenta a sobrevida de pacientes com câncer de pulmão avançado do *North Central Cancer Treatment Group* (THERNEAU, 2022).

```
#pacotes
library(mlr3)
library(mlr3viz)
library(mlr3extralearners)
library(mlr3proba)

# importando dados
lung = survival::lung
lung = na.omit(lung)
lung$status = (lung$status == 2L)

#definindo o task - dados
task = TaskSurv$new("lung",
backend = lung, time = "time",
event = "status")

#definindo os learners
learners = c(lrn("surv.rpart"),
lrn("surv.coxph"),
lrn("surv.svm",gamma.mu=0.002),
lrn("surv.coxtime"))

#definindo processo de reamostragem
resamplings = rsmp("cv", folds = 3)

# comparacao
design = benchmark_grid(task, learners, resamplings)
bmr = benchmark(design)

# resultados
bmr$aggregate()
autoplot(bmr)
```

Código R 4.6 – Exemplo - dados *Lung*

Saída do *script* R (resultado) disponível no Anexo B.

4.5 Otimização de Hiperparâmetros

Os algoritmos de aprendizado de máquina são compostos por um conjunto de hiperparâmetros, em que, a escolha dos seus respectivos valores pode impactar consideravelmente no desempenho preditivo dos modelos selecionados (BISCHL et al., 2021). Segundo Elgeldawi et al. (2021), hiperparâmetro é um parâmetro cujo valor é definido antes de iniciar o processo de aprendizado e é por meio do treinamento dos dados que os valores dos parâmetros do modelo são obtidos.

Para selecionar uma configuração de hiperparâmetro pode-se recorrer a valores padrão de hiperparâmetros que são especificados na implementação de pacotes de *software* ou configurá-los manualmente, por exemplo, com base em recomendações da literatura, experiência ou tentativa e erro. Alternativamente, pode-se usar estratégias de ajuste de hiperparâmetros, que são dependentes de dados, procedimentos de otimização de segundo nível, que tentam minimizar o erro de generalização esperado do algoritmo de indução sobre um espaço de busca de hiperparâmetros de configurações candidatas consideradas, geralmente avaliando previsões em um conjunto de teste, ou executando um esquema de reamostragem, como validação cruzada (BISCHL et al., 2012; PROBST; BOULESTEIX; BISCHL, 2019). Além disso, deve-se escolher o algoritmo de otimização. Dentre os disponíveis, destaca-se o *grid search* - busca em grade - que testa exaustivamente todas as combinações possíveis dos hiperparâmetros; e o *random search* - busca aleatória - que testa combinações aleatórias e os melhores resultados funciona como um guia para a escolha dos próximos hiperparâmetros (ELGELDAWI et al., 2021).

Os algoritmos considerados são métodos comuns de aprendizado de máquina. Os modelos a serem avaliados e seus hiperparâmetros definidos estão apresentados na Tabela 3. A escolha de hiperparâmetros e o respectivo intervalo de valores válidos (domínio) foi baseado no trabalho de (PROBST; BOULESTEIX; BISCHL, 2019)

Tabela 3 – Hiperparâmetros considerados dos algoritmos que serão otimizados

Algoritmo	**Parâmetro**	**Tipo de var.**	**Inferior**	**Superior**	**Trafo**
rpart - Árvore de Sobrevivência					
	cp	contínua	0	1	-
	maxdepth	inteira	1	30	-
	minbucket	inteira	1	60	-
	minsplit	inteira	1	60	-
Coxnet - Rede Neural de Sobrevivência					
	activation	categórica	-	-	-
	num_nodes	inteira	10	50	-
	dropout	contínua	0	1	-
	frac	contínua	0	1	-
	batch_norm	lógica (V/F)	-	-	-
svm - Máquina de Vetor de Suporte de Sobrevivência					
	kernel	categórica	-	-	-
	gamma	contínua	-10	10	2^x

Na Tabela 3, as colunas Inferior e Superior indicam o intervalo de possíveis valores dos hiperparâmetros. A função de transformação (coluna Trafo), se houver, indica como os valores são transformados. A transformação exponencial é aplicada para obter mais valores candidatos em regiões com hiperparâmetros menores porque para esses hiperparâmetros as diferenças de desempenho entre valores menores são potencialmente maiores do que para valores maiores.

Para as árvores de sobrevivência os hiperparâmetros definidos foram: *cp* - custo-complexidade, que regulariza a penalidade na função objetivo e controla a quantidade de poda; o *maxdepth* limita o crescimento da árvore além de uma certa profundidade/altura; o *minbucket* define o menor número de observações que são permitidas em um nó terminal; e o *minsplit* determina o menor número de observações no nó inicial que pode ser dividido ainda mais (THERNEAU; ATKINSON, 2022).

Em relação aos algoritmos de redes neurais de sobrevivência, definimos os seguintes hiperparâmentros: *activation*, que determina função de ativação utilizado na rede, dentre as disponíveis testamos *"celu", "elu", "gelu", "glu", "hardshrink","prelu", "rrelu", "relu", "selu", "sigmoid"*; o *num_nodes* define o número de nós presentes nas camadas ocultas; o *dropout* define a quantidade de *dropout* [2], ou seja, a porcentagem de neurônios "apagados"; a *frac* indica a fração de dados a ser usada para o conjunto de dados de validação; e o *batch_norm* determina a opção de fazer a normalização ou não em lote nas camadas (SONABEND, 2022).

Os hiperparâmetros selecionados para o algoritmo Máquina de Vetor de Suporte foram o *kernel*, que define a função de transformação dos dados, a função *svm* aceita os seguintes *kernels*: *radial, linear, polynomial* e *sigmoide*; e o parâmetro *gamma* controla a distância de influência de pontos no treinamento. Assim, quanto menor o valor de *gamma*, maior o raio de similaridade que resulta em mais pontos sendo agrupados. Enquanto que, quanto maior, menor deve ser a distância entre os pontos (observações mais similares) para serem considerados no mesmo grupo (YILDIRIM, 2020; MEYER et al., 2021).

Várias medidas são utilizadas para avaliar o desempenho dos modelos de aprendizado de máquina, entretanto, para dados censurados, a métrica tradicional é o índice C. Assim, a estimativa do índice C para os diferentes experimentos de hiperparâmetros foi obtida por meio do processo de reamostragem validação cruzada 10-*folds*.

[2] A técnica modifica o processo natural da rede, aos pouco elimina aleatoriamente (e temporariamente) alguns dos neurônios ocultos na rede.

4.5.1 Exemplo - código R para otimização de hiperparâmetros

```r
#pacotes
library(mlr3)
library(mlr3extralearners)
library(mlr3proba)
library(mlr3tuning)
library(paradox)

# importando dados
lung = survival::lung
lung = na.omit(lung)
lung$status = (lung$status == 2L)

#definindo o task - dados
task = TaskSurv$new("lung",
backend = lung, time = "time",
event = "status")

#definindo o learner e os intervalos de otimizacao
learner = lrn("surv.coxtime")
learner$param_set$values$activation = to_tune(p_fct(levels = c("celu",
    "elu", "gelu","glu","hardshrink","prelu", "rrelu", "relu", "selu", "
    sigmoid")))
learner$param_set$values$num_nodes = to_tune(p_int(lower = 10, upper =
    50))
learner$param_set$values$dropout = to_tune(p_dbl(lower = 0, upper = 1))
learner$param_set$values$frac = to_tune(p_dbl(lower = 0, upper = 1))
learner$param_set$values$batch_norm = to_tune(p_fct(levels = c(TRUE,
    FALSE)))

#definindo processo de reamostragem
resampling = rsmp("cv", folds = 5)

# otimizacao
instance = tune(
method = "random_search", #ou pode ser "grid_search"
task = task,
learner = learner,
resampling = resampling,
term_evals = 10,
batch_size = 5,
)

# resultado (performace)
instance$result
```

Código R 4.7 – Otimização de hiperparâmetros

Observação: Ao executar, deve-se ficar atento a versão do R utilizada. Alguns pacotes não foram atualizados para versões mais recentes.

5 RESULTADOS

Nesse trabalho, foi conduzido um estudo retrospectivo e observacional com a presença de pacientes com doença renal crônica, em estágio 5, e que iniciaram algum tratamento de terapia de substituição renal (TSR), isto é, diálise peritoneal (DP) ou hemodiálise (HD), no Hospital da Faculdade de Medicina de Botucatu (SP) durante o período de janeiro de 2014 a janeiro de 2019. A partir dos prontuários médicos, foram registrados a data de início da terapia de substituição renal, a terapia inicial, idade, sexo, presença de doenças de base, número de comorbidades, se houve internação, resultados de exames hematológicos e o desfecho, podendo ser o óbito, a retirada do paciente da pesquisa por falta de informações durante o seguimento, transplante renal, transferência de centro de diálise ou a sobrevivência.

De acordo com os princípios éticos básicos das diretrizes e normas regulamentadoras de pesquisa em seres humanos, este projeto de pesquisa foi submetido à Comissão de Ética em Pesquisa em Seres Humanos – protocolo 38647220.0.0000.5411 (Anexo A).

5.1 Descrição dos dados

As variáveis observadas neste estudo estão descritas na Tabela 4. Nela, também apresenta-se algumas medidas resumo, que são completadas nas Tabelas 5 e 6 . Com isso, pode-se traçar o perfil dos indivíduos pertencentes a amostra.

Dos pacientes acompanhados, aproximadamente metade iniciou a terapia renal subtitutiva (TRS) em hemodiálise (HD) e a outra metade com diálise peritoneal (DP), sendo 45% do sexo feminino, com idade média de 59,9 anos e 58,74% foram internados durante o acompanhamento. A concentração média encontradas no exame de sangue e urina inicial foram 257,43 pg/mL de paratormônio (pth), 9,98 g/dl de hemoglobina, 7,89 ml/min/1,73m^2 de clearence de creatinina e 3,41 g/dl de albumina. Em relação a presença de doenças de base, 33,22% apresentam diabetes, 10,61% doença glomerular, 5,91% doença arterial obstrutiva, 2,78% rins policísticos, 4,17% nefropatia isquêmica, 1,91% doença renal devido ao lúpus, 6,43% outras doenças de base, 15,83% a doença não foi identificada e 19,13% são hipertensos

Tabela 4 – Descrição das variáveis observadas

Variável	Descrição	Medidas Resumo
TRS	Terapia Renal Substitutiva HD = Hemodiálise DP = Diálise Peritoneal	 HD = 289 (49,74%) DP = 292 (50,26%)
Masculino	Sexo 0 = Feminino 1 = Masculino	 0 = 262 (45,09%) 1 = 319 (54,91%)
Diabetes	Paciente com diabetes 0 = Não 1 = Sim	 0 = 384 (66,78%) 1 = 191 (33,22%)
Hipertensao	Paciente hipertenso 0 = Não 1 = Sim	 0 = 465 (80,87%) 1 = 110 (19,13%)
Glomerulopatia	Paciente com doença glomerular 0 = Não 1 = Sim	 0 = 514 (89,39%) 1 = 61 (10,61%)
Obstrutiva	Paciente com doença arterial obstrutiva 0 = Não 1 = Sim	 0 = 541 (94,09%) 1 = 34 (5,91%)
Indeterminada	Paciente com doença de base indeterminada 0 = Não 1 = Sim	 0 = 484 (84,17%) 1 = 91 (15,83%)
Rins.policistico	Presença de rins policísticos 0 = Não 1 = Sim	 0 = 559 (97,22%) 1 = 16 (2,78%)
Outras	Outra doença de base 0 = Não 1 = Sim	 0 = 538 (93,57%) 1 = 37 (6,43%)
Isquemica	Paciente com nefropatia isquêmica 0 = Não 1 = Sim	 0 = 551 (95,83%) 1 = 24 (4,17%)
nefrite.lupica	Paciente com doença renal devido ao lúpus 0 = Não 1 = Sim	 0 = 564 (98,09%) 1 = 11 (1,91%)
2+COMORBIDADES	Número de comorbidades	média (desv. pad.) 2,46 (1,73)
OBITO	Variável indicadora de falha 0 = observação censurada 1 = óbito devido a doença renal crônica	 0 = 411 (70,74%) 1 = 170 (29,26%)
IDADE	Idade	média (desv. pad.) 59,9 (16,80)
SOBREVIDA	Tempo de acompanhamento até a falha ou censura (meses)	média (desv. pad.) 562,92 (521,25)
INTERNACAO	Paciente foi internado 0 = Não 1 = Sim	 0 = 222 (41,26%) 1 = 316 (58,74%)
pth	Exame paratormônio (pg/ml)	média (desv. pad.) 257,43 (277,42)
hb	Concentração de hemoglobina (g/dl)	média (desv. pad.) 9,98 (1,95)
creatinina	Concentração de Creatinina (mg/dl)	média (desv. pad.) 7,89 (6,84)
albumina	Concetração de Albumina (g/dl)	média (desv. pad.) 3,41 (0,65)

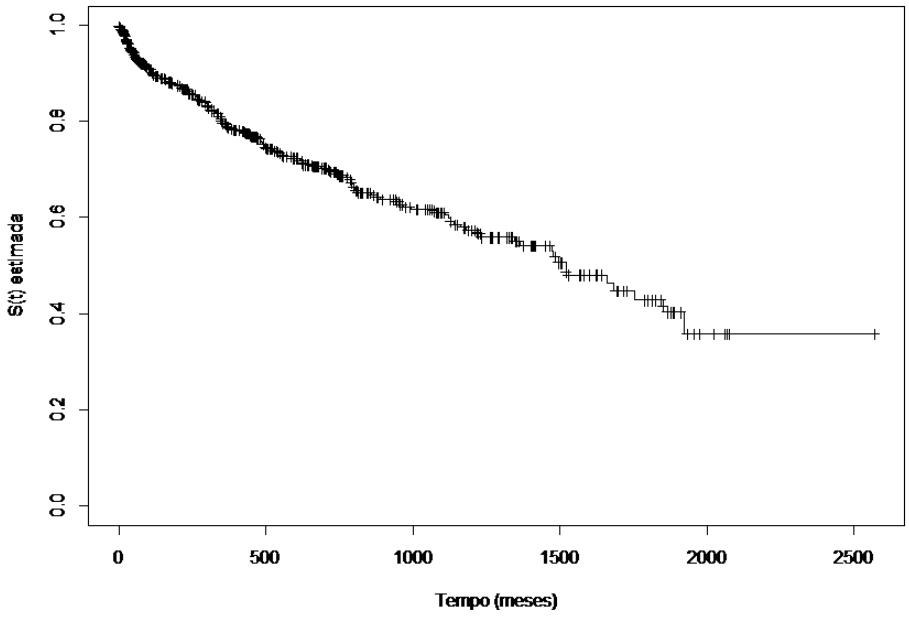
Nota: Tamanho da amostra – $n = 581$.
Obs. Para totais menores que 581, existência de dados faltantes.

Na Tabela 5, apresentam-se as frequências de algumas características observadas segundo a terapia de substituição renal iniciada nos pacientes, e também, o valor de p do teste de independência qui-quadrado ou exato de Fisher. Note que, há indícios de que o óbito ($p < 0,001$), se os pacientes apresentaram glomerulopatia ($p = 0,039$), doença de base indeterminada ($p < 0,001$), rins policísticos ($p = 0,002$) e se foram internados ($p < 0,001$), são características que diferiram entre os pacientes que iniciaram tratamento via diálise peritoneal e hemodiálise.

Enquanto que, na Tabela 6, apresentam-se algumas medidas descritivas e o valor de p do teste de comparação de médias não paramétrico de Wilcoxon, pois verificou-se que as variáveis não seguem uma distribuição normal (teste de normalidade de Shapiro-Wilk). Note que, há evidências de diferença significativa entre o número de comorbidades ($p < 0,001$) e o valor de hb ($p < 0,001$) com o tipo de tratamento iniciado.

A variável de interesse do estudo é o tempo, em meses, do início da terapia de substituição renal até o óbito. A proporção de censura nos dados é igual a 71%. Apresentam-se, na Figura 6, as estimativas da função de sobrevivência $\hat{S}(t)$ e observou-se que a curva não tende a zero. Desta forma, a função de sobrevivência é denominada imprópria.

Figura 6 – Estimativa de Kaplan-Meier da função de sobrevivência



Fonte: Produzida pela autora.

A fim de comparar as curvas de sobrevivência das categorias (grupos) das variáveis qualitativas, as funções de sobrevivência foram estimadas por grupo e comparadas pelo teste de *logrank* (Tabela 7), sob a hipótese de igualdade das curvas de sobrevivência, e representadas graficamente. Para avaliar as comparações, consideraram-se os dados por tratamento (DP ou HD) e também desconsiderando-os.

Tabela 5 – Distribuição dos pacientes segundo a terapia de substituição renal iniciada e variáveis qualitativas observadas e teste de associação. Botucatu, 2014 – 2019

Variáveis	DP		HD		p
	n	%	n	%	
Óbito					
Não	229	56	182	44	0,001 *
Sim	63	37	107	63	
Sexo					
Feminino	125	48	137	52	0,303 *
Masculino	167	52	152	48	
Diabetes					
Não	198	52	186	48	0,575 *
Sim	93	49	98	51	
Hipertensão					
Não	236	51	229	49	0,971 *
Sim	55	5	55	5	
Glomerulopatia					
Não	252	49	262	51	0,039 *
Sim	39	64	22	36	
Obstrutiva					
Não	273	50	268	50	0,917 *
Sim	18	53	16	47	
Indeterminada					
Não	224	46	260	54	0,001 *
Sim	67	74	24	26	
Rins policístico					
Não	289	52	270	48	0,002 **
Sim	2	12	14	88	
Isquêmica					
Não	282	51	269	49	0,215 **
Sim	9	38	15	63	
Nefrite Lúpica					
Não	288	51	276	49	0,138 **
Sim	3	27	8	73	
Internação					
Não	164	74	58	26	0,001 *
Sim	87	28	229	72	

Nota: * valor de p do teste qui-quadrado de associação e
** teste exato de Fisher entre a terapia de substituição renal iniciada e características clínicas observadas.

Tabela 6 – Medidas descritivas das variáveis quantitativas observadas segundo a terapia de substituição renal iniciada e teste para comparar os grupos de tratamento. Botucatu, 2014 – 2019

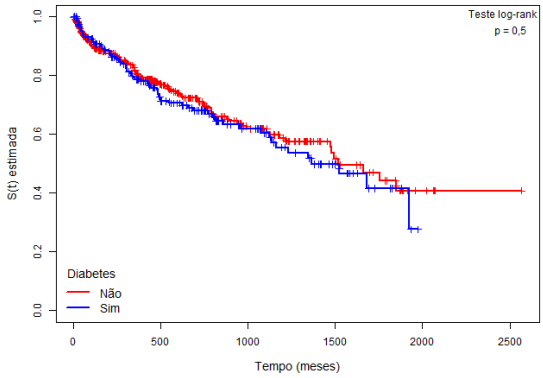
Covariáveis	Média	Desvio Padrão	Q1	Mediana	Q3	AI	p
Idade	59,9	16,8	52	62	72	20	
DP	58,99	17,19	51	62	71	20	0,286
HD	60,82	16,37	52	62	73	21	
N° de Comorbidades	2,46	1,73	1	3	4	3	
DP	1,92	1,44	1	2	3	2	0,001
HD	3	1,83	2	3	4	2	
PTH	257,43	277,42	86,65	178	319	232,35	
DP	261,63	270,19	104	182	322	218	0,165
HD	252,3	286,51	79,2	166	307	227,8	
Hemoglobina (HB)	9,98	1,95	8,6	9,9	11,2	2,6	
DP	10,29	1,95	9	10,2	11,5	2,5	0,001
HD	9,67	1,91	8,39	9,5	10,7	2,31	
Clearence creatinina	7,89	6,84	4,67	6,6	8,96	4,29	
DP	9,1	9,33	4,2	6,7	10,28	6,08	0,268
HD	6,76	2,55	4,9	6,5	8,5	3,6	
Albumina	3,41	0,65	2,98	3,5	3,9	0,93	
DP	3,37	0,64	2,9	3,45	3,8	0,9	0,215
HD	3,45	0,67	3	3,5	3,9	0,9	

Nota: p – valor de p do teste não paramétrico Wilcoxon. AI - Amplitude Interquartílica

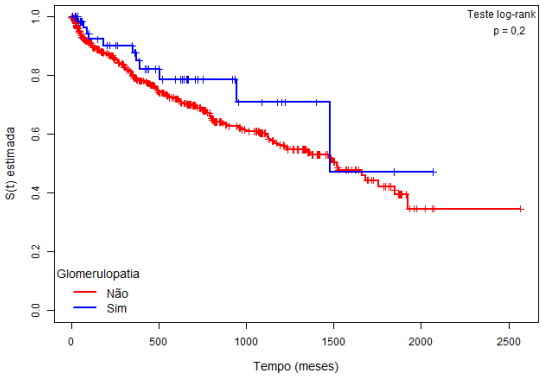
Tabela 7 – Resultados dos testes *logrank* (valor de p) para as comparações dos grupos considerando a TSR

Covariáveis	*Logrank*		
	DP+HD	DP	HD
TSR	0,2	-	-
Sexo	0,5	0,1	0,8
Diabetes	0,5	0,4	1
Hipertensão	0,8	0,9	0,7
Glomerulopatia	0,2	0,3	0,5
Obstrutiva	0,3	0,5	0,05
Indeterminada	0,5	0,5	0,2
Rins policístico	0,2	0,5	0,2
Isquêmica	0,02	0,3	0,04
Nefrite lúpica	0,4	0,4	0,6
Internação	0,002	0,001	0,3

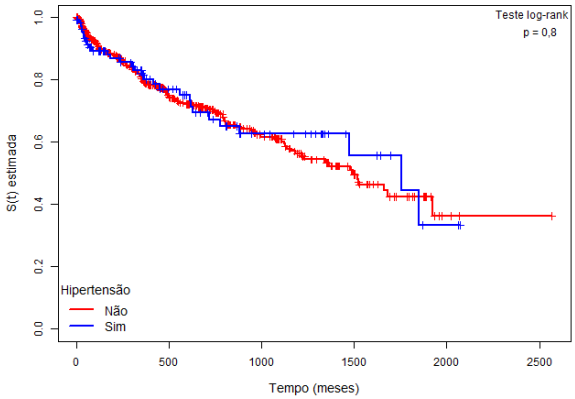
Figura 7 – Curvas de sobrevivência estimada via Kaplan-Meier das covariáveis qualitativas



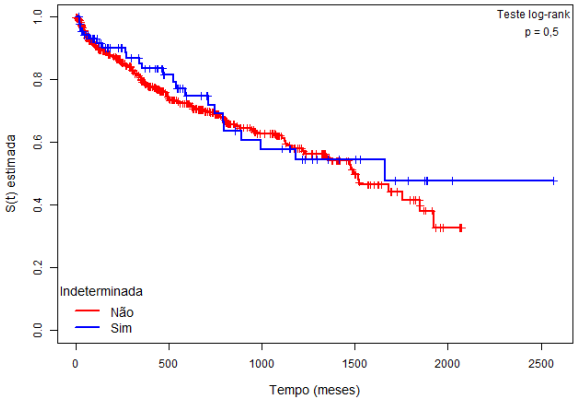
(a)



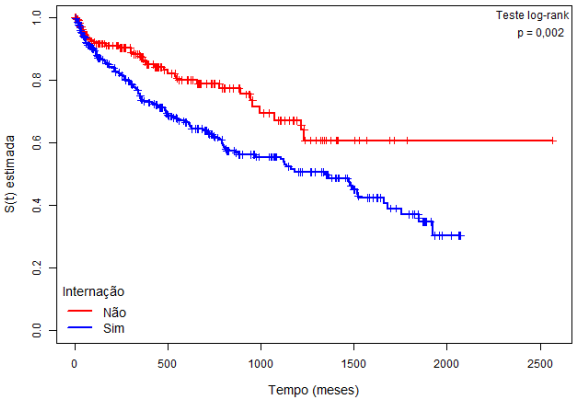
(b)



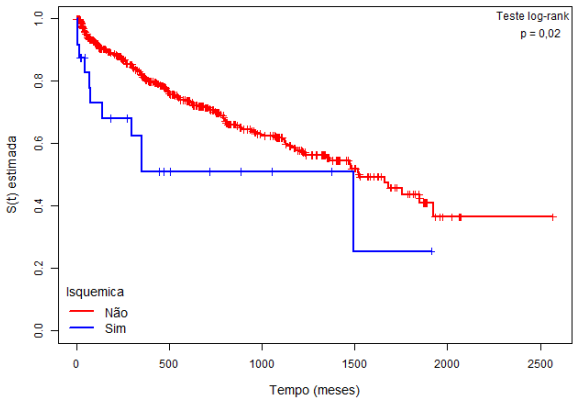
(c)



(d)



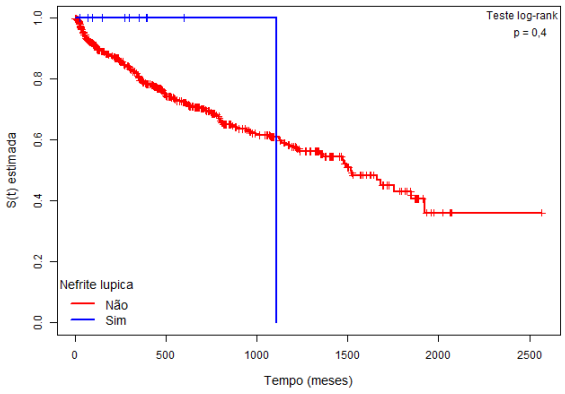
(e)



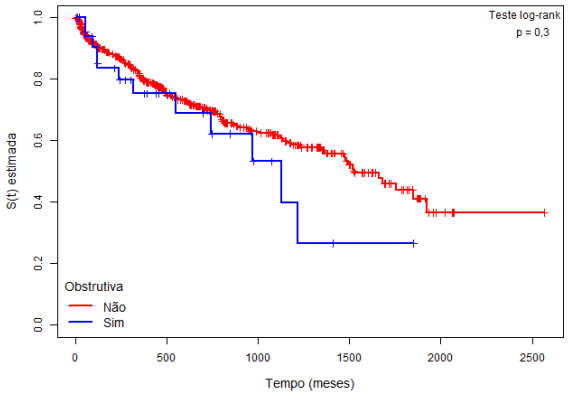
(f)

Fonte: Produzida pela autora.

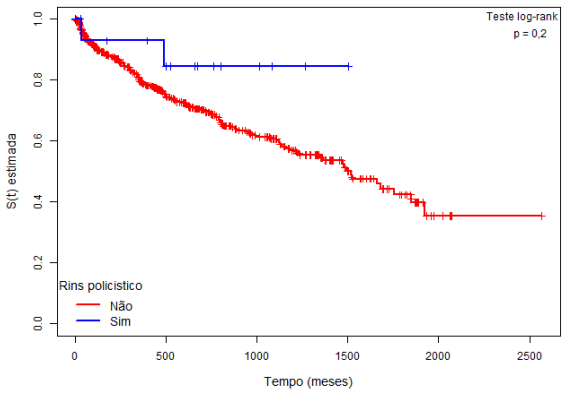
Figura 8 – Curvas de sobrevivência estimada via Kaplan-Meier das covariáveis qualitativas



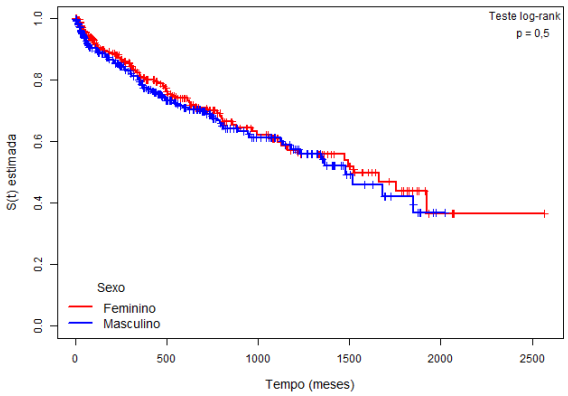
(g)



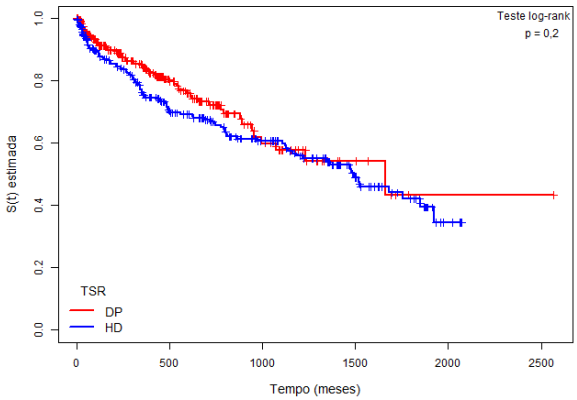
(h)



(i)



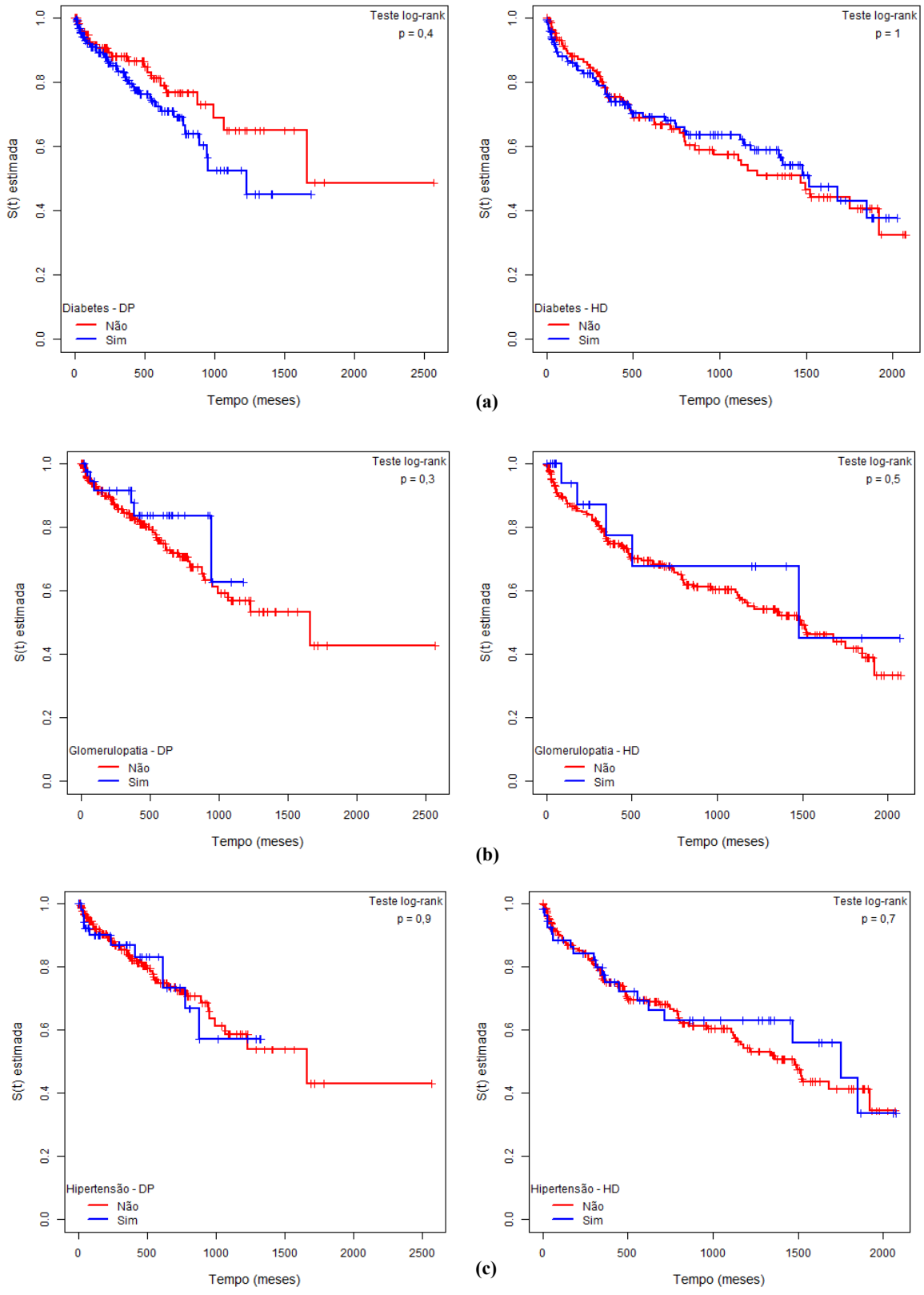
(j)



(k)

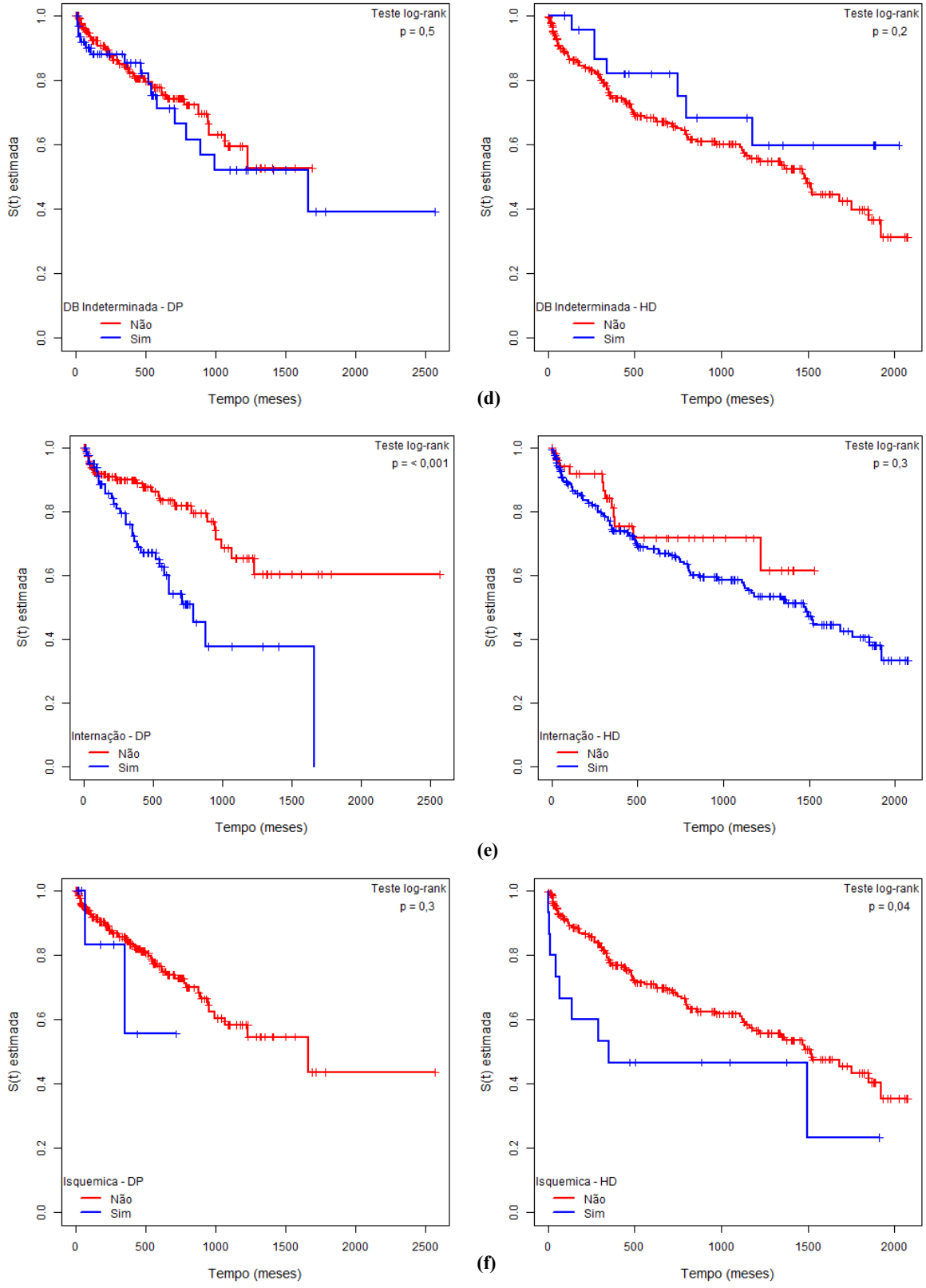
Fonte: Produzida pela autora.

Figura 9 – Curvas de sobrevivência estimada via Kaplan-Meier das covariáveis qualitativas segundo terapia de substituição renal



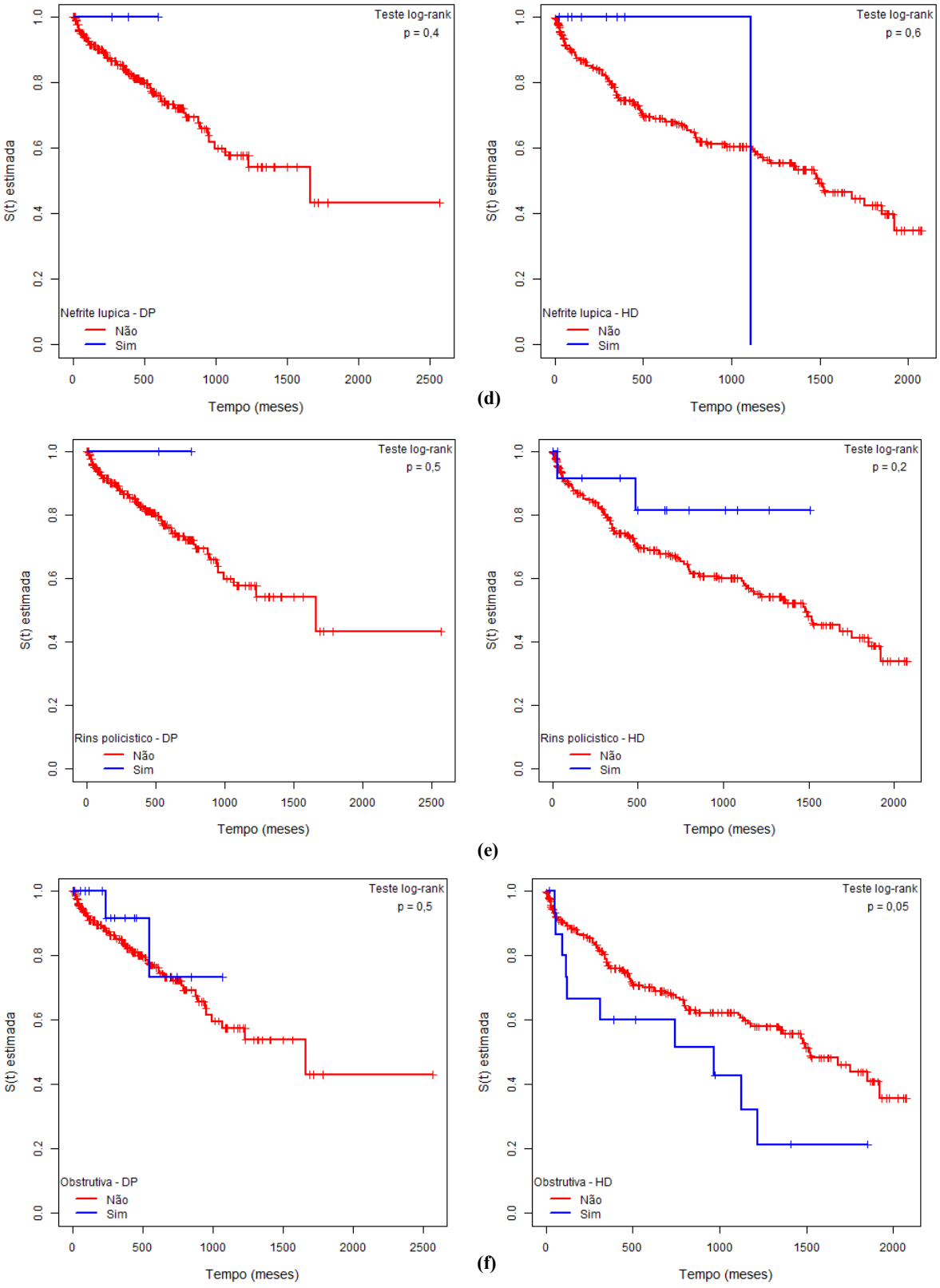
Fonte: Produzida pela autora.

Figura 10 – Curvas de sobrevivência estimada via Kaplan-Meier das covariáveis qualitativas segundo terapia de substituição renal



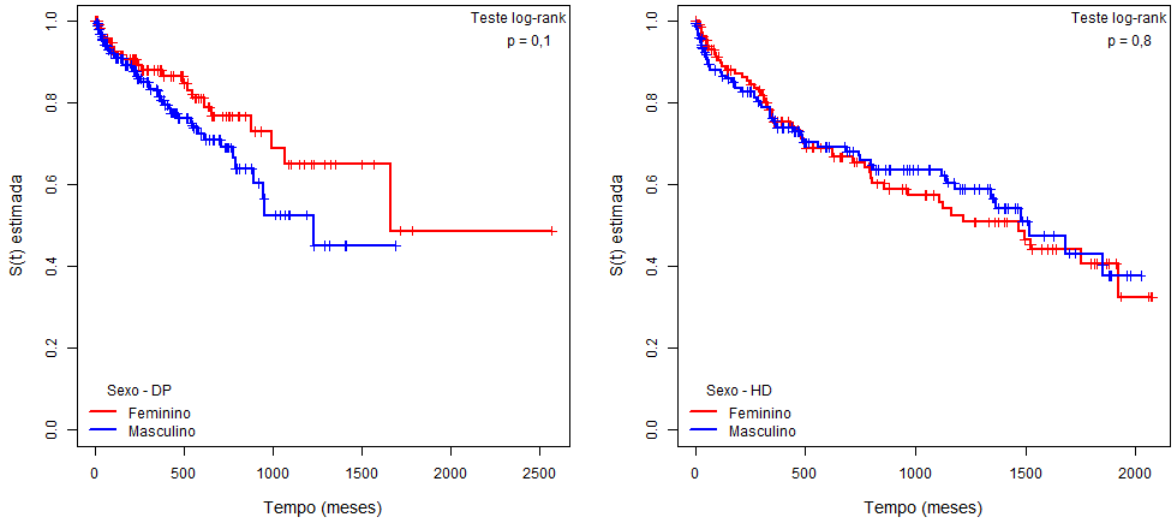
Fonte: Produzida pela autora.

Figura 11 – Curvas de sobrevivência estimada via Kaplan-Meier das covariáveis qualitativas segundo terapia de substituição renal



Fonte: Produzida pela autora.

Figura 12 – Curvas de sobrevivência estimada via Kaplan-Meier das covariáveis qualitativas segundo terapia de substituição renal



Fonte: Produzida pela autora.

Dos resultados apresentados na Tabela 7, Figura 7 e Figura 8, tem-se evidências de que existem diferenças significativas na sobrevida dentre os pacientes que apresentam ou não doença isquêmica ($p = 0,02$), e dentre os pacientes que foram e não internados ($p = 0,002$), desconsiderando o tipo de TSR iniciado. Sendo que, os pacientes expostos (isquêmicos ou internados)tiveram um tempo de sobrevivência menor dos que não expostos (Figura 7 (e) e (f)). Além disso, ao considerar os grupos de tratamento, dentre os pacientes que iniciaram o tratamento com diálise peritoneal, temos evidências de que aqueles que foram internados tiveram menor sobrevida comparado aos que não foram internaram ($p < 0,001$). Em relação aos pacientes que iniciaram o tratamento com hemodiálise, temos indícios de que os com doença obstrutiva ($p = 0,05$) ou isquêmica ($p = 0,04$) apresentaram menor tempo de sobrevivência comparados aos que não apresentaram (Figuras 9, 10, 11 e 12).

5.2 Resultados inferenciais

O modelo de Cox foi ajustado para análise preliminar e inferencial dos dados obtidos na pesquisa. Diversos modelos foram ajustados a fim de obter aquele com a melhor qualidade do ajuste, comparando-os pelo Critério de Informação de Akaike (AIC).

Na Tabela 8, apresenta-se as estimativas do modelo de Cox ajustado selecionado.

Para avaliar a adequação do modelo de Cox utilizamos os resíduos padronizados de Schoenfeld, apresentados na Tabela 9 e na Figura 13, podem ser observado que a tendência ao longo do tempo não indica forte desproporcionalidade. O teste de proporcionalidade indicou haver indícios de que as variáveis concentração de albumina e hemoglobina não apresentam

Tabela 8 – Estimativas obtidas para o modelo de Cox ajustado aos dado de pacientes renais

Covariável	**Estimativa**	**Erro Pad.**	**RR**	**1/RR**	**z**	$IC(RR)_{95\%}$	**valor p**
Idade	0,03	0,01	1,03	0,97	4,48	(1,017; 1,043)	<0,001
TRS_HD	1,32	0,51	3,73	0,27	2,61	(1,386; 10,036)	0,009
Internado_ Sim	0,75	0,29	2,12	0,47	2,61	(1,207; 3,738)	0,009
Clearence de creatinina	0,03	0,01	1,03	0,97	1,95	(1,000; 1,059)	0,051
Albumina	-0,33	0,15	0,72	1,38	-2,19	(0,539; 0,967)	0,029
Hemoglobina	-0,09	0,05	0,91	1,10	-1,88	(0,827; 1,004)	0,061
TRS_HD*Internado_Sim	-0,85	0,43	0,43	2,35	-2,00	(0,184; 0,983)	0,046
TRS_HD*Clearence.Creatinina	-0,13	0,05	0,87	1,14	-2,54	(0,789; 0,969)	0,011

Nota: 118 observações perdidas (*missing*), $n = 463$. Risco Relativo $= RR = exp(estimativa)$. $AIC = 1474,198$.

Tabela 9 – Teste da proporcionalidade dos riscos do modelo de Cox ajustado

Covariável	χ^2	**gl**	**valor p**
Idade	0,05	1	0,826
TRS_HD	3,28	1	0,070
Internado_ Sim	0,57	1	0,451
Creatinina	3,09	1	0,079
Albumina	4,30	1	0,038
Hemoglobina	4,10	1	0,043
TRS_HD*Internado_Sim	2,11	1	0,146
TRS_HD*Clearence.Creatinina	0,04	1	0,842
GLOBAL	11,85	8	0,158

riscos proporcionais. Entretanto, diferentemente para o modelo global (valor p = 0,158), por isso, optou-se em continuar com a interpretação do modelo.

Portanto, o modelo de Cox obtido é dado por:

$$\lambda(t) = \lambda_0(t) \exp\{\mathbf{X}'\boldsymbol{\beta}\}$$

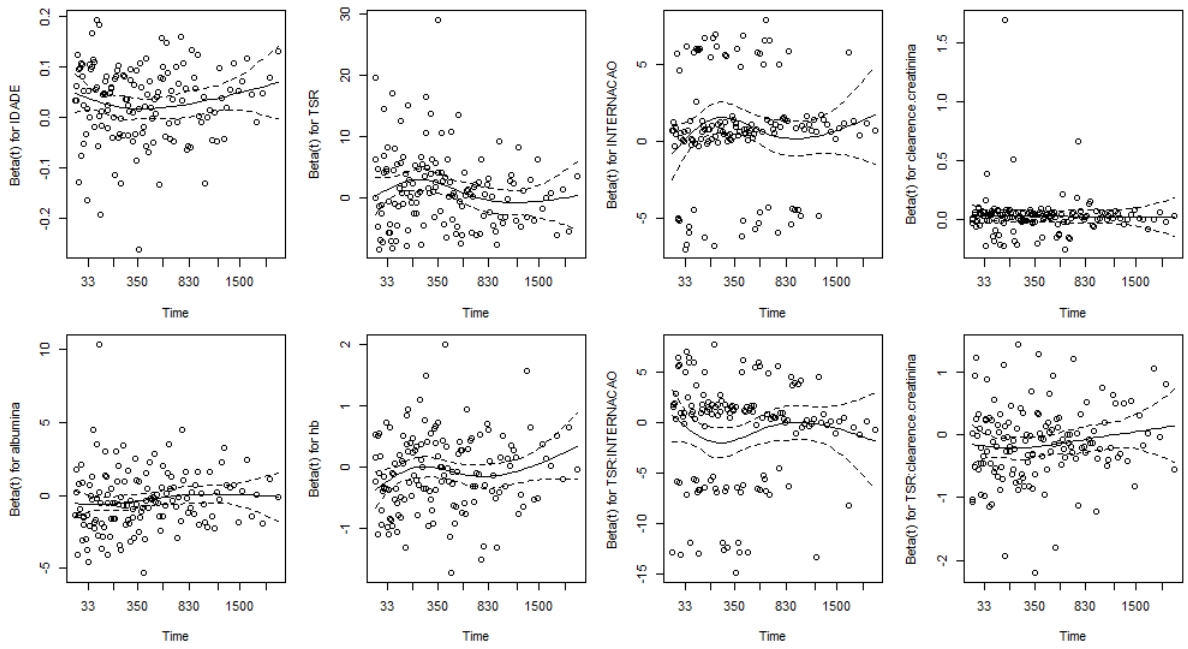
$$= \lambda_0(t) \exp\{0,03 * Idade - 0,33 * Albumina - 0,09 * Hemoglobina$$

$$+1,32 * TSR_{HD} + 0,75 * Internação + 0,03 * Creatinina$$

$$-0,85 * TSR_{HD} * Internação - 0,13 * TSR_{HD} * Creatinina\}$$

Desta forma, tem-se que o aumento em uma unidade na idade, na concentração de *clearence* de creatinina e na concentração de albumina do paciente aumenta o risco de óbito em 3% ($\exp(0,03)$) e 3% ($\exp(0,03)$), para os dois primeiros, respectivamente, e diminui o risco em 38% ($1/RR = 1/0,72 = 1,38$), para o terceiro; o risco de óbito de um paciente que iniciou a TSR com hemodiálise é 3,73 ($\exp(1,32)$) vezes o risco de pacientes que iniciaram com diálise peritoneal; um paciente que foi internado tem 2,12 ($\exp(0,75)$) vezes mais risco de óbito do que aqueles que não foram internados; o risco de óbito de um paciente que iniciou o tratamento com diálise peritoneal e foi internado é 2,35 ($1/RR = 1/0,43 = 2,35$) vezes maior que o risco de

Figura 13 – Resíduos padronizados de Schoenfeld do modelo de Cox ajustado



Fonte: Produzida pela autora.

um paciente que iniciou tratamento com hemodiálise e foi internado; ao passo que, o aumento em uma unidade na concentração de clearence de creatinina em pacientes que iniciaram o tratamento com diálise peritoneal acarreta em um aumento 14% ($1/RR = 1/0,87 = 1,14$) do risco de morte do que aqueles que iniciaram o tratamento com hemodiálise.

5.3 Resultados preditivos

Na primeira etapa, a modelagem foi realizada no conjunto de dados completo ($n = 463$) com a finalidade de selecionar as variáveis relevantes. Os diversos modelos de Cox (pacote R: *survival*) ajustados foram comparados utilizando o AIC. Em seguida, os modelos foram ajustados com os valores dos parâmetros obtidos na otimização de alguns hiperparâmetros dos algoritmos de aprendizagem de máquina: árvore de sobrevivência (pacote R: *rpart*), redes neurais de sobrevivência (pacote R: *survivalmodels*) e máquina de vetores de suporte de sobrevivência (pacote R: *survivalsvm*). Cada modelo foi avaliado estimando o índice C médio via métodos de reamostragem, validação cruzada *k-folds* e *bootstrap*.

Na Tabela 10, apresentam-se os resultados e o índice C médio para os hiperparâmetros otimizados e, também, considerando os respectivos valores padrões dos algoritmos. Observe que, o índice C, isto é, o desempenho dos modelos treinados foi melhor quando utilizado os valores otimizados dos hiperparâmetros.

As estimativas de desempenho dos modelos comparados são apresentados na Tabela 11, com base no índice C (*C-index*). Dos quatro modelos, o modelo mais adequado recebeu a

Tabela 10 – Resultados dos algoritmos otimizados e índice C médio dos modelos considerando os valores otimizados e padrões do pacote

Algoritmo	Parâmetro	Valor Otimizado	Índice C	Valor Padrão	Índice C
Árvore de Sobrevivência					
	cp	0,016	0,626	0,01	0,598
	minsplit	24		20	
	maxdepth	14		30	
	minbucket	12		7	
MVS de sobrevivência					
	gamma.mu	$2^{-}7,52$	0,675	1/p	0,365
	kernel	lin_kernel		radial	
RN de Sobrevivência (Coxtime)					
	activation	rrelu	0,614	relu	0,602
	num_nodes	26		32	
	dropout	0,039		NULL	
	frac	0,280		0	
	batch_norm	FALSE		TRUE	

Nota: p é número de variáveis do conjunto de dados.

classificação 1, o menos, recebeu a classificação 4. O modelo Cox teve desempenho preditivo mais adequado dentre todos os modelos de sobrevivência treinados. Enquanto que, o modelo de rede neural (*coxtime*) teve o desempenho menos adequado do que os demais modelos. É possível observar resultados semelhantes nas Figuras 14 e 15 em relação ao poder de predição dos modelos. Nota-se também que, as estimativas obtidas via bootstrap foram mais precisas, enquanto que, por meio da metodologia de validação cruzada, as estimativas foram mais dispersas.

Tabela 11 – Desempenho da previsão dos modelos de sobrevivência

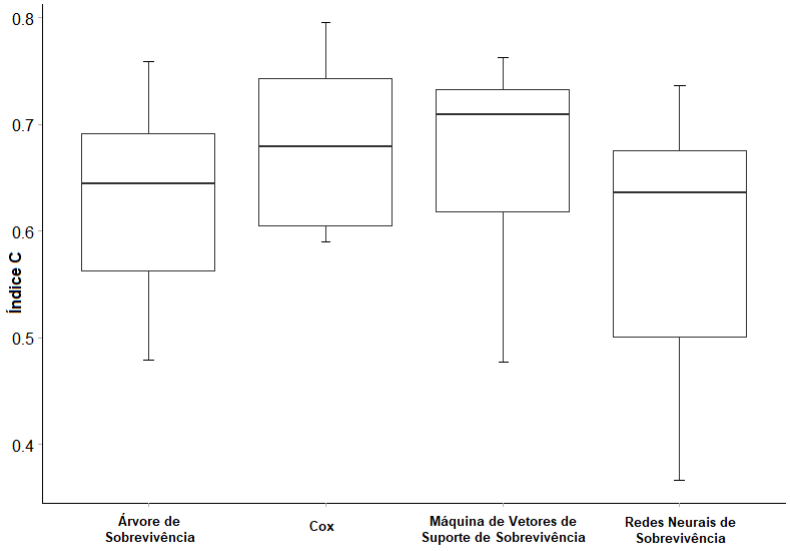
Modelo	CV 10-*fold*	*Bootstrap*	Classificação
Máquina de Vetor de Suporte de Sobrevivência	0,663	0,653	2
Rede Neural de Sobrevivência	0,589	0,583	4
Cox - Riscos Proporcionais	0,680	0,689	1
Árvore de Sobrevivência	0,629	0,609	3

Nota: Os valores internos são as médias dos índices C (*C-index*) obtidos nas amostras.
Bootstrap ($n = 10$)

De uma perspectiva preditiva, na Tabela 12, para cada cenário, apresenta-se a previsão do risco de óbito do paciente em relação a um padrão que utiliza a média do valor para cada covariável como referência. Enquanto que, na Figura 16, apresentam-se as previsões das curvas de sobrevivência para cada cenário, obtidas a partir do modelo de Cox treinado.

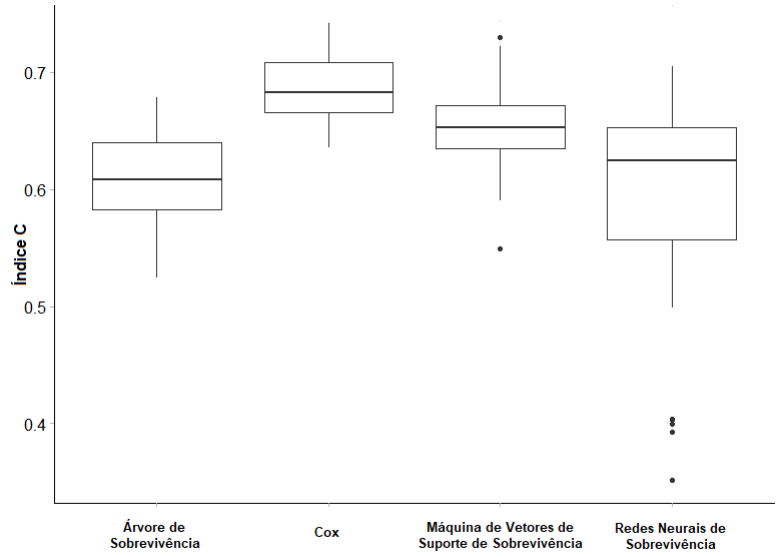
Comparando os cenários (Tabela 12), os riscos e as curvas de sobrevivência preditas (Figura 16), nota-se que pacientes próximos ao Cenário 3 são os que possuem maior risco de óbito. Percebe-se ainda, que o fator determinante para o aumento do risco é a internação dos

Figura 14 – Box plot dos índices C das reamostragens - validação cruzada 10 *folds* - para os diferentes modelos de sobrevivência



Fonte: Produzida pela autora.

Figura 15 – Box plot dos índices C das reamostragens - *Bootstrap* - para os diferentes modelos de sobrevivência



Fonte: Produzida pela autora.

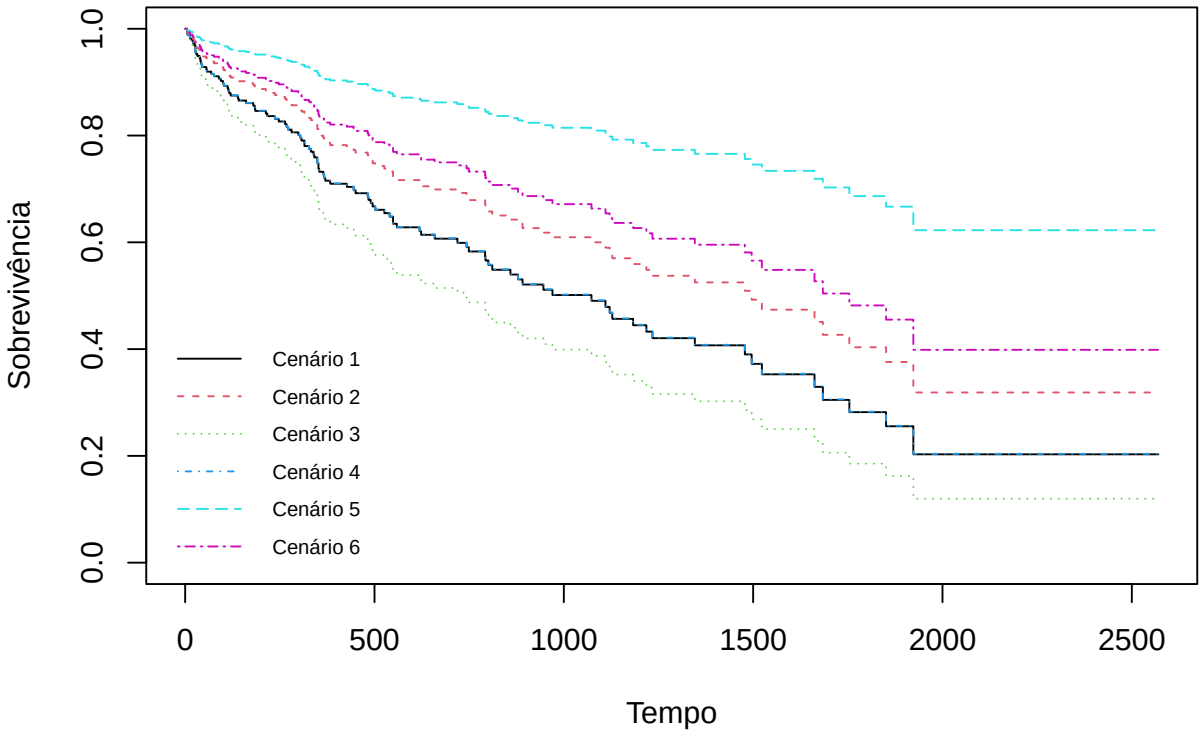
pacientes, seguido por aqueles que iniciaram o tratamento com a diálise peritoneal e o aumento da idade.

Tabela 12 – Previsão de risco de óbito de pacientes segundo alguns cenários

Cenário	Idade	Albumina	Hemoglobina	Hemodiálise	Internação	Creatinina	RR
1	56	3,3	8,8	0	1	6,49	3,88
2	66	4,3	8,3	1	1	6,3	2,78
3	71	3,5	8,3	0	1	5	5,16
4	89	3,8	10	0	0	14,7	3,87
5	50	3,4	7,7	1	0	12,7	1,15
6	64	2,6	10,6	0	0	6,93	2,23

Nota: $RR = \exp\{\mathbf{x}_i'\boldsymbol{\beta}\}$

Figura 16 – Previsão da estimativa da sobrevivência segundo cenário



Fonte: Produzida pela autora.

6 DISCUSSÃO E CONSIDERAÇÕES FINAIS

Tradicionalmente, modelos de aprendizado de máquina são úteis para lidar com problemas complexos incluindo, nos últimos anos, grande base de dados. Neste trabalho, o objetivo foi obter modelos de predição, em especial, modelos de aprendizado de máquina para dados censurados. Para avaliar o poder preditivo, utilizou-se o índice C. A análise descritiva permitiu identificar as possíveis variáveis significativas para explicar o tempo de sobrevivência dos pacientes em terapia renal substitutiva. Além da análise descritiva, também foi realizada uma seleção de covariáveis a partir dos modelos de regressão de Cox ajustados com a finalidade de obter modelos mais precisos, corroborando com os resultados obtidos por Spooner et al. (2020).

Trabalhos como de Maccariello et al. (2008), Segal et al. (2020), Kim et al. (2021) e Monaghan et al. (2021) também optaram por obter modelos preditivos para pacientes renais focados em objetivos distintos, porém centrados em modelos de classificação, em que a variável resposta é uma variável qualitativa. No entanto, abordagens para dados de sobrevivência com esta aplicação, como os abordados aqui, ainda são pouco exploradas.

Com este estudo, observou-se haver indícios de que a sobrevivência dos pacientes que iniciaram a terapia renal substitutiva (TRS) em diálise peritoneal (DP) é maior do que aqueles que iniciaram com hemodiálise (HD). Entretanto, pacientes que foram internados e iniciaram o tratamento em HD, terão menor risco de morte do que aqueles que iniciaram com DP. Além disso, desconsiderando o tratamento iniciado, pacientes que forem internados têm maior risco de morte. A idade também foi significativa como fator de risco, como indicado em Termorshuizen et al. (2003), Bastos et al. (2009), Pereira et al. (2016). González et al. (2015) observaram que a sobrevivência é maior entre os pacientes submetidos à hemodiálise do que à diálise peritoneal e também que, aqueles pacientes que iniciaram o tratamento com a diálise peritoneal e depois mudaram para hemodiálise obtiveram maior sobrevivência do que aqueles que apenas foram tratados por meio da hemodiálise.

Alguns trabalhos mostraram melhores resultados com a DP no grupo de pacientes jovens e sem comorbidades (KOREVAAR et al., 2003; SESSO et al., 2012). Entretanto, outros estudos evidenciaram menor mortalidade após 2 anos de terapia dialítica em pacientes idosos e com

comorbidades, dentre elas: insuficiência cardíaca, insuficiência coronariana e diabetes mellitus (TERMORSHUIZEN et al., 2003; VONESH et al., 2006). Enquanto que, neste estudo, nenhuma doença de base foi significativa para explicar a sobrevivência dos pacientes em tratamento dialítico, apesar de ser fator de risco.

Kvamme, Borgan e Scheel (2019) propuseram extensões do modelo de riscos proporcionais de Cox (*Cox-time*) e compararam com regressão de Cox clássica, florestas aleatória de sobrevivência e modelos de *deep learning* para dados de sobrevivência. Diferente dos nossos resultados, o método *Cox-Time* teve o melhor desempenho geral, ao passo que, o modelo de *deep learning* (*DeepHit*) teve o melhor desempenho discriminativo, custando as estimativas de sobrevivência. Enquanto que, Hazewinkel, Gelderblom e Fiocco (2022) compararam os modelos obtidos por meio dos riscos proporcionais de Cox, redes neurais de sobrevivência e florestas aleatória de sobrevivência aplicados ao estudo clínico de tumor maligno ósseo (osteossarcoma) e observaram que os três métodos tiveram desempenho semelhante.

As limitações encontradas no trabalho foram a grande quantidade de censura (aproximadamente 71%) dos tempos de sobrevivência, a complexidade do conjunto de dados aplicados aos métodos de aprendizado de máquina tradicionalmente utilizados, a dificuldade em encontrar pacotes do *software* R atualizados para análise do problema e o desafio de buscar exemplos de programação da metodologia de aprendizado de máquina com a inclusão de observações censuradas. Ademais, observou-se que a função de sobrevivência é imprópria, indicativo da presença de fração de cura ou de sobreviventes de longa duração, no entanto, o pouco explorado indicou que o modelo de Cox ainda teve um ajuste melhor (via AIC). Considerando este resultado, explorou-se também, a comparação de modelos de sobrevivência para dados que não atendesse a suposição de riscos proporcionais, entretanto, as predições não foram boas e o modelo de Cox também teve desempenho mais preciso (via índice C).

Portanto, visando as limitações descritas, um estudo que pode ser explorado futuramente é a comparação entre os diferentes métodos de redes neurais que melhor exploram as camadas ocultas, por exemplo o *deep learning*, como apresentado no artigo de Kvamme, Borgan e Scheel (2019), pois, Katzman et al. (2018b) concluíram que o uso de *deep learning* na análise de sobrevivência permite maior desempenho devido à flexibilidade do modelo e a metodologia prevê o risco de forma equivalente ou melhor do que outros métodos de sobrevivência lineares e não lineares.

REFERÊNCIAS

BASTOS, M. G.; KIRSZTAJN, G. M. Doença renal crônica: importância do diagnóstico precoce, encaminhamento imediato e abordagem interdisciplinar estruturada para melhora do desfecho em pacientes ainda não submetidos à diálise. *J Bras Nefrol*, v. 33, n. 1, p. 93–108, 2011. 3, 4

BASTOS, R. M. R. et al. Prevalência da doença renal crônica nos estágios 3, 4 e 5 em adultos. *Revista da Associação Médica Brasileira*, SciELO Brasil, v. 55, p. 40–44, 2009. 52

BECKER, M. et al. mlr3 book. *URl: https://mlr3book. mlr-org. com*, 2021. 28, 29, 30

BISCHL, B. et al. Hyperparameter optimization: Foundations, algorithms, best practices and open challenges. *arXiv preprint arXiv:2107.05847*, 2021. 32

BISCHL, B. et al. mlr: Machine learning in r. *Journal of Machine Learning Research*, v. 17, n. 170, p. 1–5, 2016. Disponível em: <https://jmlr.org/papers/v17/15-066.html>. 28

BISCHL, B. et al. Resampling methods for meta-model validation with recommendations for evolutionary computation. *Evolutionary computation*, MIT Press, v. 20, n. 2, p. 249–275, 2012. 32

BRASIL. *Diretrizes clínicas para o cuidado ao paciente com doença renal crônica – DRC no Sistema Único de Saúde*. Brasília: Ministério da Saúde - Secretaria de Atenção à Saúde - Departamento de Atenção Especializada e Temática, 2014. 3, 4

BRASIL. *Doenças Renais Crônicas (DRC)*. [S.l.]: Ministério da Saúde, 2021. <https://www.gov.br/saude/pt-br/assuntos/saude-de-a-a-z/d/doencas-renais>. Acesso em 11/2021. 4

BRASIL. *Saúde apresenta atual cenário das doenças não transmissíveis no Brasil*. Brasília: Ministério da Saúde, 2021. <https://www.gov.br/saude/pt-br/assuntos/noticias/2021-1/setembro/saude-apresenta-atual-cenario-das-doencas-nao-transmissiveis-no-brasil>. Acesso em 10/10/21. 1

BRASILIA. *Doença renal crônica é epidêmica, diz Sociedade Brasileira de Nefrologia*. Senado Federal: Secretaria de Comunicação Social: Agência Senado, 2020. <https://senado.jusbrasil.com.br/noticias/820456222/doenca-renal-cronica-e-epidemica-diz-sociedade-brasileira-de-nefrologia>. Acesso em 15/06/22. 1

BREIMAN, L. Random forests. *Machine learning*, Springer, v. 45, n. 1, p. 5–32, 2001. 14, 15

BREIMAN, L. et al. *Classification and Regression Tree*. [S.l.]: Chapman and Hall, 1984. 18, 20, 21

BRESLOW, N. Contribuição à discussão do artigo de dr cox. *Journal of the Royal Statistical Society, Series B*, v. 34, p. 216–7, 1972. 12

BURGER, S. V. *Introduction to machine learning with R: Rigorous mathematical analysis*. [S.l.]: O'Reilly Media, Inc., 2018. 14

CHAUDHARY, K.; SANGHA, H.; KHANNA, R. Peritoneal dialysis first: Rationale. *Clin J Am Soc Nephrol*, v. 6, p. 447—-456, 2011. 6

CHEN, L. Overview of clinical prediction models. *Annals of translational medicine*, AME Publications, v. 8, n. 4, 2020. 2

CIAMPI, A. et al. Stratification by stepwise regression, correspondence analysis and recursive partition: a comparison of three methods of analysis for survival data with covariates. *Computational Statistics & Data Analysis*, v. 4, n. 3, p. 185–204, 1986. 20

COLLETT, D. *Modelling Survival Data in Medical Research*. 2nd. ed. Florida: Chapman and Hall/CRC, 2003. 391 p. 1, 2, 7, 9, 10

COLOSIMO, E. A.; GIOLO, S. R. *Análise de sobrevivência aplicada*. São Paulo: Blucher, 2006. 370 p. 7, 8, 9, 10, 11, 12, 13

CORTES, C.; VAPNIK, V. Support-vector networks. *Machine learning*, v. 20, n. 3, p. 273–297, 1995. 24

COX, D. R. Partial likelihood. *Biometrika*, Oxford University Press, v. 62, n. 2, p. 269–276, 1975. 11

DIAS, D. B. et al. Peritoneal dialysis as an option for unplanned initiation of chronic dialysis. *Hemodialysis International*, v. 20, n. 4, p. 631–633, 2016. 4, 5, 6

DIAS, D. B. et al. Urgent-start peritoneal dialysis: The first year of brazilian experience. *Blood Purif*, v. 44, p. 283—-287, 2017. 5, 6

ELGELDAWI, E. et al. Hyperparameter tuning for machine learning algorithms used for arabic sentiment analysis. In: MDPI. *Informatics*. [S.l.], 2021. v. 8, n. 4, p. 79. 32

FARAGGI, D.; SIMON, R. A neural network model for survival data. *Statistics in medicine*, Wiley Online Library, v. 14, n. 1, p. 73–82, 1995. 24

FERREIRA, E. V. Métodos de reamostragem. *Material de apoio a disciplina de Machine Learning para Cientista de Dados, lecionada na LEG/UFPR*, 2018. 16, 17

FOUODO, C. J. et al. Support vector machines for survival analysis with r. *R Journal*, v. 10, n. 1, 2018. 25, 26, 27

FRIEDMAN, J. et al. *The elements of statistical learning*. [S.l.]: Springer series in statistics New York, 2001. v. 1. 21, 23, 24, 26

GALLETTI, A. J. d. F. *Modelos de fração de cura aplicados aos tempos de sobrevivência de pacientes submetidos à ligadura elástica de varizes no esôfago*. Dissertação (Mestrado em Biometria) — Universidade Estadual Paulista (UNESP), Botucatu, SP, 2018. 9

GEHAN, E. A. A generalized wilcoxon test for comparing arbitrarily singly-censored samples. *Biometrika*, v. 52, n. 1/2, p. 203–223, 1965. 9

GERRISH, S. *How smart machines think*. [S.l.]: MIT Press, 2018. 21

GOLLAPUDI, S. *Practical machine learning*. [S.l.]: Packt Publishing Ltd, 2016. 2, 14, 15, 18

GONZáLEZ, A. O. et al. Supervivencia en hemodiálisis vs. diálisis peritoneal y por transferencia de técnica. experiencia en ourense 1976-2012. *Revista de la Sociedad Española de Nefrología*, v. 35, n. 6, p. 562—-566, 2015. 6, 52

GORDON, L.; OLSHEN, R. A. Tree-structured survival analysis. *Cancer Treatment Reports*, p. 1065–1068, 1985. 20

HARRELL, F. E. et al. Evaluating the yield of medical tests. *Jama*, American Medical Association, v. 247, n. 18, p. 2543–2546, 1982. 17

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2ed. ed. New York, USA: Springer, 2017. 745 p. (Series in Statistics). 2

HAYKIN, S. *Neural networks and learning machines, 3/E*. New Jersey: Pearson Education, 2009. 906 p. 22, 23

HAZEWINKEL, A.-D.; GELDERBLOM, H.; FIOCCO, M. Prediction models with survival data: a comparison between machine learning and the cox proportional hazards model. *medRxiv*, Cold Spring Harbor Laboratory Press, 2022. 53

HECHANOVA, L. A. *Diálise*. [S.l.]: Texas Tech University Health Sciences Center, 2020. <https://www.msdmanuals.com/pt-pt/casa/multimedia/image/KID_hemodialysis_peritoneal_pt>. Acesso em 06/2022. 5

IBGE. *Pesquisa Nacional de Saúde 2013: Percepção do estado de saúde, estilos de vida e doenças crônicas*. Rio de Janeiro: Instituto Brasileiro de Geografia e Estatística - IBGE, 2014. 1

IVARSEN, P.; POVLSEN, J. V. Can peritoneal dialysis be applied for unplanned initiation of chronic dialysis? *Nephrol Dial Transplant*, v. 29, p. 2201—-2206, 2014. 5, 6

IZBICKI, R.; SANTOS, T. M. dos. *Aprendizado de máquina: uma abordagem estatística*. [S.l.]: Rafael Izbicki, 2020. 15, 22, 23, 24, 25

JAMES, G. et al. *An introduction to statistical learning*. [S.l.]: Springer, 2013. v. 112. 19

JIN, H.; LU, Y. Cost-saving tree-structured survival analysis for hip fracture of study of osteoporotic fractures data. *Medical Decision Making*, SAGE Publications Sage CA: Los Angeles, CA, v. 31, n. 2, p. 299–307, 2011. 20

JIN, H. et al. Alternative tree-structured survival analysis based on variance of survival time. *Medical Decision Making*, v. 24, n. 6, p. 670–680, 2004. 20

JOHNSON, R. J.; FEEHALLY, J.; FLOEGE, J. *Nefrologia clínica: abordagem abrangente*. [S.l.]: Elsevier Brasil, 2016. 4, 5

KALBFLEISCH, J. D.; PRENTICE, R. L. *The Statistical Analysis of Failure Time Data*. 2. ed. New Jersey: John Wiley & Sons, 2002. 439 p. (Wiley Series in Probability and Statistics). 9

KAPLAN, E. L.; MEIER, P. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, v. 53, n. 282, p. 457–481, 1958. 8

KATZMAN, J. L. et al. Deepsurv: personalized treatment recommender system using a cox proportional hazards deep neural network. *BMC medical research methodology*, BioMed Central, v. 18, n. 1, p. 1–12, 2018. 24

KATZMAN, J. L. et al. Deepsurv: personalized treatment recommender system using a cox proportional hazards deep neural network. *BMC medical research methodology*, BioMed Central, v. 18, n. 1, p. 1–12, 2018. 53

KIM, H. W. et al. Dialysis adequacy predictions using a machine learning method. *Scientific reports*, Nature Publishing Group, v. 11, n. 1, p. 1–7, 2021. 2, 52

KLEINBAUM, D. G.; KLEIN, M. *Survival Analysis: A Self-Learning Text*. 3. ed. New York: Springer-Verlag, 2012. 700 p. (Statistics for Biology and Health). 1, 2, 7, 11

KOREVAAR, J. C. et al. Effect of starting with hemodialysis compared with peritoneal dialysis in patients new on dialysis treatment: a randomized controlled trial. *Kidney Int*, v. 64, n. 6, p. 2222–2228, 2003. 52

KUHN, M.; JOHNSON, K. *Feature engineering and selection: A practical approach for predictive models*. [S.l.]: CRC Press, 2019. 16

KVAMME, H.; BORGAN, Ø.; SCHEEL, I. Time-to-event prediction with neural networks and cox regression. *arXiv preprint arXiv:1907.00825*, 2019. 53

LANG, M. et al. mlr3: A modern object-oriented machine learning framework in R. *Journal of Open Source Software*, dec 2019. Disponível em: <https://joss.theoj.org/papers/10.21105/joss.01903>. 28

LAWLESS, J. F. *Statistical models and methods for lifetime data*. [S.l.]: John Wiley & Sons, 2011. 11

LEBLANC, M.; CROWLEY, J. Relative risk trees for censored survival data. *Biometrics*, JSTOR, p. 411–425, 1992. 20, 21

LEE, Y.; BANG, H.; KIM, D. J. How to establish clinical prediction models. *Endocrinol Metab*, v. 31, p. 38–44, 2016. 14

LUCAS, L. d. S. Árvores, florestas e sua função como preditores: Uma aplicação na avaliação do grau de maturidade de empresas. *Rev. Pmkt*, v. 4, n. 1, p. 6–11, 2011. 18

MACCARIELLO, E. R. et al. Desempenho de seis modelos de predição prognóstica em pacientes críticos que receberam suporte renal extracorpóreo. *Revista Brasileira de Terapia Intensiva*, SciELO Brasil, v. 20, p. 115–123, 2008. 2, 52

MANTEL, N. Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemotherapy Rep.*, v. 50, p. 163–170, 1966. 9, 10

MARINHO, A. W. G. B. et al. Prevalência de doença renal crônica em adultos no brasil: revisão sistemática da literatura. *Cad. Saúde Colet.*, v. 25, n. 3, p. 379–388, 2017. 1, 3, 6

MAYER, F. P. Métodos de reamostragem. *Material de apoio a disciplina de Estatística Computacional II, lecionada na LEG/UFPR*, 2021. 16, 17

MENDES, M. L. et al. Peritoneal dialysis as the first dialysis treatment option initially unplanned. *Brazilian Journal of Nephrology*, SciELO Brasil, v. 39, p. 441–446, 2017. 5, 6

MEYER, D. et al. *e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien*. [S.l.], 2021. R package version 1.7-9. Disponível em: <https://CRAN.R-project.org/package=e1071>. 33

MOISEN, G. Classification and regression trees. *In: Jørgensen, Sven Erik; Fath, Brian D.(Editor-in-Chief). Encyclopedia of Ecology, volume 1. Oxford, UK: Elsevier. p. 582-588.*, p. 582–588, 2008. 19

MONAGHAN, C. K. et al. Machine learning for prediction of patients on hemodialysis with an undetected sars-cov-2 infection. *Kidney360*, American Society of Nephrology, v. 2, n. 3, p. 456, 2021. 2, 52

MORGAN, J. N.; SONQUIST, J. A. Problems in the analysis of survey data, and a proposal. *Journal of the American statistical association*, Taylor & Francis, v. 58, n. 302, p. 415–434, 1963. 18

MUELLER, J. P.; MASSARON, L. *Machine learning for dummies*. [S.l.]: John Wiley & Sons, 2021. 15, 25

NERBASS, F. B. et al. Censo brasileiro de diálise 2020. *Brazilian Journal of Nephrology*, SciELO Brasil, 2022. 1

PEREIRA, E. R. S. et al. Prevalence of chronic renal disease in adults attended by the family health strategy. *Brazilian Journal of Nephrology*, SciELO Brasil, v. 38, p. 22–30, 2016. 52

PONCE, D.; BRABOA, A. M.; BALBIA, A. L. Urgent start peritoneal dialysis. *Wolters Kluwer Health*, v. 27, 2018. 5, 6

PROBST, P.; BOULESTEIX, A.-L.; BISCHL, B. Tunability: importance of hyperparameters of machine learning algorithms. *The Journal of Machine Learning Research*, JMLR. org, v. 20, n. 1, p. 1934–1965, 2019. 32

R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2021. Disponível em: <https://www.R-project.org/>. 28

ROJAS, R. *Neural networks: a systematic introduction*. [S.l.]: Springer Science & Business Media, 2013. 23

SBN. *Diálise Peritoneal*. [S.l.]: Sociedade Brasileira de Nefrologia, 2021. <https://www.sbn.org.br/orientacoes-e-tratamentos/tratamentos/dialise-peritoneal/>. Acesso em 10/2021. 4

SBN. *Hemodiálise*. [S.l.]: Sociedade Brasileira de Nefrologia, 2021. <https://www.sbn.org.br/orientacoes-e-tratamentos/tratamentos/hemodialise/>. Acesso em 10/2021. 4

SBN. *Compreendendo os rins*. [S.l.]: Sociedade Brasileira de Nefrologia, 2022. <https://www.sbn.org.br/o-que-e-nefrologia/compreendendo-os-rins/>. Acesso em 05/2022. 3

SBN. *Doença renal crônica: diagnóstico e prevenção*. [S.l.]: Sociedade Brasileira de Nefrologia, 2022. <https://www.sbn.org.br/noticias/single/news/doenca-renal-cronica-diagnostico-e-prevencao/>. Acesso em 05/2022. 4

SCHMID, M.; WRIGHT, M. N.; ZIEGLER, A. On the use of harrell's c for clinical risk prediction via random survival forests. *Expert Systems with Applications*, Elsevier, v. 63, p. 450–459, 2016. 17

SEGAL, Z. et al. Machine learning algorithm for early detection of end-stage renal disease. *BMC nephrology*, Springer, v. 21, n. 1, p. 1–10, 2020. 2, 52

SESSO, R. C. C. et al. Diálise crônica no brasil - relatório do censo brasileiro de diálise. *J Bras Nefrol*, v. 34, n. 3, p. 272–277, 2012. 52

SHIVASWAMY, P. K.; CHU, W.; JANSCHE, M. A support vector approach to censored targets. In: IEEE. *Seventh IEEE international conference on data mining (ICDM 2007)*. [S.l.], 2007. p. 655–660. 26, 27

SILVA, A. R. et al. Doenças crônicas não transmissíveis e fatores sociodemográficos associados a sintomas de depressão em idosos. *J Bras Psiquiatr.*, v. 66, n. 1, p. 45–51, 2017. 1

SONABEND, R. *survivalmodels: Models for Survival Analysis*. [S.l.], 2022. R package version 0.1.11. Disponível em: <https://CRAN.R-project.org/package=survivalmodels>. 33

SONABEND, R. et al. mlr3proba: An r package for machine learning in survival analysis. *Bioinformatics*, 02 2021. ISSN 1367-4803. 28, 29

SPOONER, A. et al. A comparison of machine learning methods for survival analysis of high-dimensional clinical data for dementia prediction. *Scientific reports*, Nature Publishing Group, v. 10, n. 1, p. 1–10, 2020. 2, 52

STEYERBERG, E. W. *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*. New York: Springer, 2009. 497 p. 2, 14

TACONELI, C. *Árvores de classificação multivariadas fundamentadas em coeficientes de dissimilaridade e entropia100 p*. 100 p. Tese (Tese de Doutorado) — Tese (Doutorado em Estatística e Experimentação Agronômica)–Escola Superior de Agricultura Luiz de Queiroz, Piracicaba, SP, 2008. 18, 19

TAY, K. *What is Harrell's C-index?* 2019. <https://statisticaloddsandends.wordpress.com/2019/10/26/what-is-harrells-c-index/>. Acesso em 06/2022. 17

TERMORSHUIZEN, F. et al. Hemodialysis and peritoneal dialysis: comparison of adjusted mortality rates according to the duration of dialysis: analysis of the netherlands cooperative study on the adequacy of dialysis. *J Am Soc Nephrol.*, v. 14, n. 11, p. 2851–2860, 2003. 52, 53

THERNEAU, T.; ATKINSON, B. *rpart: Recursive Partitioning and Regression Trees*. [S.l.], 2022. R package version 4.1.16. Disponível em: <https://CRAN.R-project.org/package=rpart>. 33

THERNEAU, T. M. *A Package for Survival Analysis in R*. [S.l.], 2022. R package version 3.3-1. Disponível em: <https://CRAN.R-project.org/package=survival>. 31

THERNEAU, T. M.; GRAMBSCH, P. M. *Modeling Survival Data: Extending the Cox Model*. 1. ed. [S.l.]: Springer, 2000. (Statistics for Biology and Health). ISBN 9781441931610; 1441931619; 9781475732948; 1475732945. 13

THERNEAU, T. M.; GRAMBSCH, P. M.; FLEMING, T. R. Martingale-based residuals for survival models. *Biometrika*, v. 77, n. 1, p. 147–160, 1990. 20

VAPNIK, V. *The nature of statistical learning theory*. [S.l.]: Springer science & business media, 1999. 24, 27

VAPNIK, V.; LERNER, A. Y. Recognition of patterns with help of generalized portraits. *Avtomat. i Telemekh*, v. 24, n. 6, p. 774–780, 1963. 24

VONESH, E. F. et al. Mortality studies comparing peritoneal dialysis and hemodialysis: What do they tell us? *Kidney International*, v. 70, p. S3—-S11, 2006. 53

WANG, P.; LI, Y.; REDDY, C. K. Machine learning for survival analysis: A survey. *ACM Computing Surveys (CSUR)*, ACM New York, NY, USA, v. 51, n. 6, p. 1–36, 2019. 2, 20

YILDIRIM, S. Hyperparameter tuning for support vector machines-c and gamma parameters. *URL: https://towardsdatascience. com/hyperparameter-tuning-for-support-vectormachines-c-and-gamma-parameters-6a5097416167*, 2020. 33

Anexos

A Parecer da Comissão de Ética em Pesquisa em Seres Humanos

PARECER CONSUBSTANCIADO DO CEP

DADOS DO PROJETO DE PESQUISA

Título da Pesquisa: Modelos de predição para dados censurados aplicados a pacientes com doença renal crônica em terapia de substituição renal

Pesquisador: Agda Jessica de Freitas Galletti

Área Temática:

Versão: 2

CAAE: 38647220.0.0000.5411

Instituição Proponente: Faculdade de Medicina de Botucatu/UNESP

Patrocinador Principal: Financiamento Próprio

DADOS DO PARECER

Número do Parecer: 4.521.605

Apresentação do Projeto:

Será um estudo retrospectivo, observacional que aplicará diferentes algoritmos de predição para dados censurados em doença renal crônica, visando identificar fatores de risco e/ou de proteção que interfiram na sobrevivência de pacientes em tratamento dialítico. Serão coletados dados sociodemográficos e clínicos, a partir de 2012, de 1000 pacientes portadores de doença renal crônica atendidos no Serviço de Diálise do Hospital da Faculdade de Medicina de Botucatu, Botucatu, SP.

Objetivo da Pesquisa:

Identificar fatores de risco e/ou de proteção que possam interferir na sobrevivência de pacientes com doença renal em tratamento dialítico utilizando diferentes algoritmos de predição para dados censurados.

Avaliação dos Riscos e Benefícios:

Os riscos são mínimos, desde que seja garantida a manutenção do sigilo e da privacidade dos participantes. Os benefícios são indiretos, visando que, futuramente, os achados auxiliem no tratamento da doença.

Comentários e Considerações sobre a Pesquisa:

Trata-se de um projeto de pesquisa relevante e factível, retrospectivo, com dados a partir de 2012. não informa até que data será o uso dos dados. O custo será de R$100,00 com

financiamento próprio.

os pesquisadores responderam as pendencias apontadas pelo colegiado:

PENDÊNCIA 1. Esclarecer qual período de uso dos dados para o estudo (de 2012 a 2020?)
RESPOSTA: Serão incluídos pacientes que iniciaram a terapia renal substitutiva de janeiro de 2014 a janeiro de 2019. Portanto, o período de uso dos dados será de 05 anos (janeiro de 20014 a janeiro de 2019). O período da coleta de dados será de janeiro a março de 2021.

PENDÊNCIA 2. Apresentar critérios de inclusão e exclusão
RESPOSTA: Pacientes com doença renal crônica e que iniciaram algum tratamento de terapia de substituição renal, isto é, diálise peritoneal ou hemodiálise, durante o período da pesquisa será o critério para incluí-los no estudo. Durante este período predefinido, serão registrados a data de início da terapia de substituição renal, das internações, das complicações infecciosas, da troca de técnica de tratamento (diálise peritoneal para hemodiálise) e do desfecho, podendo ser o óbito, a retirada do paciente da pesquisa por falta de informações durante o seguimento, transplante renal, transferência de centro de diálise, ou a sobrevivência. Não há critérios de exclusão.

PENDÊNCIA 3. Esclarecer como a principal variável de interesse será o tempo de seguimento até o óbito - serão incluídos apenas no estudo pacientes que foram à óbito? Se sim, justificar adequadamente a dispensa do TCLE. Se não, isto é, se o estudo incluir pacientes que estão em seguimento no serviço - apresentar TCLE para estes participantes.
RESPOSTA: A variável de interesse do estudo é o tempo, em meses, do início da terapia de substituição renal até o óbito. Entretanto, a informação dos pacientes que sobreviveram por todo o período de acompanhamento é muito importante pois além de identificar os fatores que influenciaram no óbito, pode identificar também os fatores que influenciaram na longevidade destes pacientes pós início da terapia de substituição renal. Apresentado TCLE.

Considerações sobre os Termos de apresentação obrigatória:
Apresentam-se anuências institucionais: Serviço de Diálise do HCFMB, Anuência HCFMB e Anuência FMB. Os pesquisadores apresentaram TCLE para aplicação aos pacientes em seguimento no serviço.

Continuação do Parecer: 4.521.605

Recomendações:

Apresentar relatório final de atividades após encerramento da pesquisa.

A coleta de dados deverá ser iniciada após aprovação do CEP.

Conclusões ou Pendências e Lista de Inadequações:

Após análise em REUNIÃO ORDINÁRIA, o Colegiado deliberou APROVADO o projeto de pesquisa apresentado.

Considerações Finais a critério do CEP:

Conforme deliberação do Colegiado, em REUNIÃO ORDINÁRIA do Comitê de Ética em Pesquisa FMB/UNESP, realizada em 01/02/2021, o Projeto de Pesquisa encontra-se APROVADO.

A coleta de dados deverá ser iniciada após data de aprovação do CEP.

Apresentar relatório final de atividades após finalização da pesquisa.

Att.

CEP-FMB

Este parecer foi elaborado baseado nos documentos abaixo relacionados:

Tipo Documento	Arquivo	Postagem	Autor	Situação
Informações Básicas do Projeto	PB_INFORMAÇÕES_BÁSICAS_DO_PROJETO_1337776.pdf	25/11/2020 22:05:48		Aceito
Outros	carta_resposta.doc	25/11/2020 22:00:43	Agda Jessica de Freitas Galletti	Aceito
TCLE / Termos de Assentimento / Justificativa de Ausência	TCLE.docx	25/11/2020 21:57:05	Agda Jessica de Freitas Galletti	Aceito
Projeto Detalhado / Brochura Investigador	projeto_Galletti_Agda_v2.pdf	25/11/2020 21:55:32	Agda Jessica de Freitas Galletti	Aceito
Declaração de Instituição e Infraestrutura	AnuenciaChefeDaDialise.pdf	25/09/2020 00:48:39	Agda Jessica de Freitas Galletti	Aceito
Declaração de Instituição e Infraestrutura	TermoDeAnuenciaInstitucional.pdf	25/09/2020 00:46:07	Agda Jessica de Freitas Galletti	Aceito
Declaração de Instituição e Infraestrutura	AnuenciaHcfmbSipe3442020.pdf	25/09/2020 00:45:15	Agda Jessica de Freitas Galletti	Aceito
Folha de Rosto	FolhaDeRostoAssinada.pdf	25/09/2020	Agda Jessica de	Aceito

Continuação do Parecer: 4.521.605

Folha de Rosto	FolhaDeRostoAssinada.pdf	00:38:58	Freitas Galletti	Aceito

Situação do Parecer:
Aprovado

Necessita Apreciação da CONEP:
Não

BOTUCATU, 03 de Fevereiro de 2021

Assinado por:
SILVANA ANDREA MOLINA LIMA
(Coordenador(a))

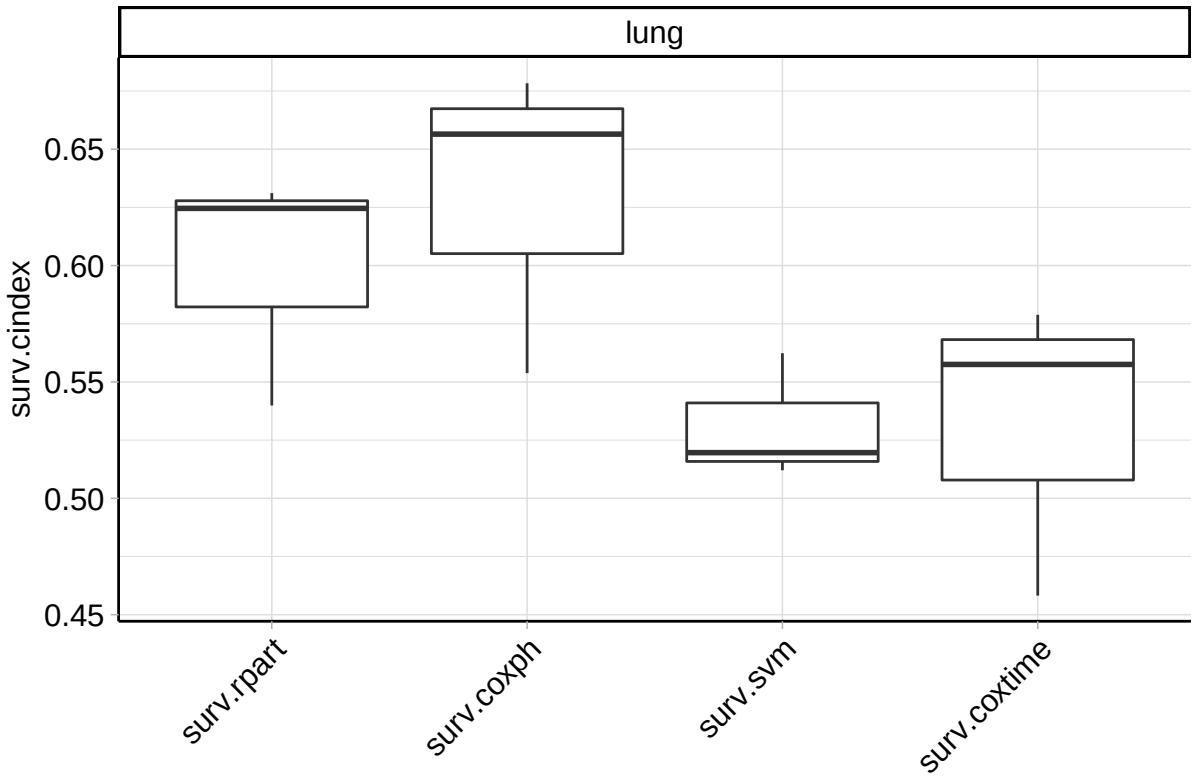
Endereço: Chácara Butignolli , s/n
Bairro: Rubião Junior **CEP:** 18.618-970
UF: SP **Município:** BOTUCATU
Telefone: (14)3880-1609 **E-mail:** cep@fmb.unesp.br

B Saída (*output*) - exemplo - dados lung

Output - exemplo - dados lung

```
## INFO  [02:58:23.232] [mlr3] Running benchmark with 12 resampling iterations
## INFO  [02:58:23.390] [mlr3] Applying learner 'surv.rpart' on task 'lung' (iter 1/3)
## INFO  [02:58:23.515] [mlr3] Applying learner 'surv.rpart' on task 'lung' (iter 2/3)
## INFO  [02:58:23.552] [mlr3] Applying learner 'surv.rpart' on task 'lung' (iter 3/3)
## INFO  [02:58:23.581] [mlr3] Applying learner 'surv.coxph' on task 'lung' (iter 1/3)
## INFO  [02:58:23.650] [mlr3] Applying learner 'surv.coxph' on task 'lung' (iter 2/3)
## INFO  [02:58:23.681] [mlr3] Applying learner 'surv.coxph' on task 'lung' (iter 3/3)
## INFO  [02:58:23.722] [mlr3] Applying learner 'surv.svm' on task 'lung' (iter 1/3)
## INFO  [02:58:24.617] [mlr3] Applying learner 'surv.svm' on task 'lung' (iter 2/3)
## INFO  [02:58:24.950] [mlr3] Applying learner 'surv.svm' on task 'lung' (iter 3/3)
## INFO  [02:58:25.592] [mlr3] Applying learner 'surv.coxtime' on task 'lung' (iter 1/3)
## INFO  [02:59:10.103] [mlr3] Applying learner 'surv.coxtime' on task 'lung' (iter 2/3)
## INFO  [02:59:10.467] [mlr3] Applying learner 'surv.coxtime' on task 'lung' (iter 3/3)
## INFO  [02:59:10.806] [mlr3] Finished benchmark
```

```
##    nr      resample_result task_id   learner_id resampling_id iters surv.cindex
## 1:  1 <ResampleResult[21]>    lung   surv.rpart            cv     3   0.5985321
## 2:  2 <ResampleResult[21]>    lung   surv.coxph            cv     3   0.6295357
## 3:  3 <ResampleResult[21]>    lung     surv.svm            cv     3   0.5313444
## 4:  4 <ResampleResult[21]>    lung surv.coxtime            cv     3   0.5315210
```


lung
surv.cindex
0.65
0.60
0.55
0.50
0.45
surv.rpart
surv.coxph
surv.svm
surv.coxtime