Universidade Estadual de Maringá
Departamento de Estatística
Programa de Pós-graduação em Bioestatística

Juliana Nascimento de Paula

DIFERENCIAÇÃO DE BACTEREMIAS GRAM-POSITIVAS E GRAM-NEGATIVAS ATRAVÉS DE BIOMARCADORES LABORATORIAIS DE ROTINA: ESTABILIDADE DA SELEÇÃO DE VARIÁVEIS E MODELOS PREDITIVOS

Maringá
2026

Juliana Nascimento de Paula

DIFERENCIAÇÃO DE BACTEREMIAS GRAM-POSITIVAS E GRAM-NEGATIVAS ATRAVÉS DE BIOMARCADORES LABORATORIAIS DE ROTINA: ESTABILIDADE DA SELEÇÃO DE VARIÁVEIS E MODELOS PREDITIVOS

Dissertação apresentada ao Programa de Pós-graduação em Bioestatística do Centro de Ciências Exatas da Universidade Estadual de Maringá, como requisito parcial para a obtenção do título de Mestre em Bioestatística.

Orientador: Prof. Dr. Josmar Mazucheli

Universidade Estadual de Maringá - UEM

Maringá
2026

Dados Internacionais de Catalogação-na-Publicação (CIP)
(Biblioteca Central - UEM, Maringá - PR, Brasil)

P324d

Paula, Juliana Nascimento de

Diferenciação de bacteremias Gram-positivas e Gram-negativas através de biomarcadores laboratoriais de rotina : estabilidade da seleção de variáveis e modelos preditivos / Juliana Nascimento de Paula. -- Maringá, PR, 2026.

117 f. : il. color., figs., tabs.

Orientador: Prof. Dr. Josmar Mazucheli.

Dissertação (mestrado) - Universidade Estadual de Maringá, Centro de Ciências Exatas, Departamento de Estatística, Programa de Pós-Graduação em Bioestatística, 2026.

1. Bacteremia. 2. Biomarcadores. 3. Infecções bacterianas. 4. Bioestatística. 5. Métodos estatísticos. I. Mazucheli, Josmar, orient. II. Universidade Estadual de Maringá. Centro de Ciências Exatas. Departamento de Estatística. Programa de Pós-Graduação em Bioestatística. III. Título.

CDD 23.ed. 519.5

Rosana de Souza Costa de Oliveira - 9/1366

JULIANA NASCIMENTO DE PAULA

Diferenciação de Bacteremias Gram-Positivas e Gram-Negativas Através de Biomarcadores Laboratoriais de Rotina: Estabilidade da Seleção de Variáveis e Modelos Preditivos

Dissertação apresentada ao Programa de Pós-Graduação em Bioestatística do Centro de Ciências Exatas da Universidade Estadual de Maringá, como requisito parcial para a obtenção do título de Mestre em Bioestatística.

BANCA EXAMINADORA

Prof. Dr. Josmar Mazucheli
Universidade Estadual de Maringá – PBE/UEM

Prof. Dr. Vanderly Janeiro
Universidade Estadual de Maringá – PBE/UEM

Dra. Daniela Dambroso Altafini
Hospital Universitário – HUM/UEM

Maringá, 26 de fevereiro de 2026.

Agradecimentos

Agradeço, em primeiro lugar, ao professor Dr. Josmar Mazucheli por me confiar este tema, pela orientação e pelas sugestões e contribuições que possibilitaram a realização deste trabalho, bem como por favorecer meu crescimento e desenvolvimento acadêmico.

Aos membros da banca Dra. Daniela Dambroso Altafini e Prof. Dr. Vanderly Janeiro, minha sincera gratidão pela disponibilidade e pelas valiosas contribuições desde a etapa de qualificação. Agradeço, em especial, à Dra Daniela, pela disponibilização dos dados que tornaram este trabalho possível, e pelas considerações que certamente o enriquecerão.

Agradeço ao Programa de Pós-Graduação em Bioestatística e aos professores que contribuíram para minha formação ao longo do mestrado. Ao secretário do programa, Daniel, minha gratidão pela atenção, disponibilidade e auxílio em todas as questões administrativas.

Agradeço à CAPES pelo apoio financeiro, por meio da concessão de bolsa de mestrado, que tornou possível a dedicação a esta pesquisa.

Aos meus colegas e amigos da salinha, Jessica, Pedro, Márcio, Ícaro, Dayana, Haward e João, obrigada pela parceria, pelas conversas, pela troca de ideias e pelo apoio nos momentos mais desafiadores desta caminhada.

Às amigas do Palafito, Clara, Eloisa, Dany, Bárbara e Raquel, pelo carinho, pela leveza e pela presença constante ao longo desta caminhada. Vocês são especiais.

Aos meus pais, Isaias a Sonia, meu tio Eduardo, agradeço com todo carinho pelo amor, pela compreensão e pelo apoio constante, que me deram força para seguir em frente. Àqueles que mesmo de longe demonstram seu amor e torcida por mim, Leonardo, Priscila e Isadora. E, em especial, à Mariana, que além de irmã, é uma amiga, incentivadora e inspiração para mim, muito obrigada, de coração.

Por fim, agradeço imensamente a Deus pelas pessoas que Ele coloca em meu caminho, por me sustentar ao longo dessa trajetória, concedendo sabedoria, serenidade e perseverança para chegar até aqui.

"O que sabemos é uma gota; o que ignoramos é um oceano"
Isaac Newton

Resumo

A diferenciação precoce entre bacteremias Gram-positivas e Gram-negativas constitui um desafio clínico relevante, especialmente nas primeiras horas de atendimento, quando decisões terapêuticas são tomadas antes do resultado das hemoculturas. Neste trabalho, técnicas de aprendizado estatístico associadas à seleção estável de variáveis foram aplicadas a dados laboratoriais de rotina com o objetivo de ajudar a discriminar a etiologia bacteriana de infecções da corrente sanguínea. A análise utilizou dados previamente coletados de 523 pacientes atendidos em hospital universitário, incluindo biomarcadores laboratoriais e variáveis clínicas disponíveis no momento da coleta da hemocultura. A estabilidade da seleção de preditores foi avaliada por meio da regularização LASSO combinada a procedimentos de reamostragem, especificamente *bootstrap* e *subsampling*, em comparação com métodos sequenciais clássicos. Esse procedimento permitiu identificar conjuntos de variáveis com diferentes níveis de parcimônia e robustez, posteriormente utilizados na construção e avaliação de modelos preditivos por regressão logística, máquina de suporte e modelos em conjunto. Os resultados evidenciaram que a qualidade e a estabilidade do conjunto de preditores exercem maior influência sobre o desempenho discriminatório do que a complexidade do algoritmo de classificação. O conjunto estável ampliado apresentou melhor discriminação global, enquanto o modelo parcimonioso baseado em Creatinina, Idade, Lactato e Bilirrubina Total demonstrou desempenho competitivo e elevada sensibilidade, indicando potencial aplicabilidade como ferramenta de triagem inicial. De forma consistente, marcadores associados à disfunção orgânica e à resposta inflamatória sistêmica concentraram a maior relevância preditiva, sugerindo que exames laboratoriais de rotina contêm informação suficiente para apoiar a inferência etiológica precoce em bacteremias.

Palavras-chaves: bacteremia; biomarcadores laboratoriais de rotina; seleção de variáveis baseada em estabilidade; modelagem preditiva; regressão logística.

Abstract

Early differentiation between Gram-positive and Gram-negative bacteremia constitutes a relevant clinical challenge, particularly during the first hours of care, when therapeutic decisions are made before blood culture results become available. In this study, statistical learning techniques combined with stability-based variable selection were applied to routine laboratory data in order to help discriminate the bacterial etiology of bloodstream infections. The analysis used previously collected data from 523 patients treated at a university hospital, including laboratory biomarkers and clinical variables available at the time of blood sample collection. The stability of predictor selection was evaluated through LASSO regularization combined with resampling procedures, specifically bootstrap and subsampling, in comparison with classical sequential methods. This procedure allowed the identification of sets of variables with different levels of parsimony and robustness, which were subsequently used in the construction and evaluation of predictive models using logistic regression, support vector machines, and ensemble models. The results showed that the quality and stability of the predictor set exert greater influence on discriminative performance than the complexity of the classification algorithm. The expanded stable set presented better overall discrimination, while the parsimonious model based on Creatinine, Age, Lactate, and Total Bilirubin demonstrated competitive performance and high sensitivity, indicating potential applicability as an initial screening tool. Consistently, markers associated with organ dysfunction and systemic inflammatory response concentrated the greatest predictive relevance, suggesting that routine laboratory tests contain sufficient information to support early etiological inference in bacteremia.

Key-words: bacteremia; routine laboratory biomarkers; stability selection; predictive modeling; logistic regression.

Lista de ilustrações

Lista de tabelas

Sumário

Capítulo 1

Introdução

A presença de bactérias na corrente sanguínea é denominada bacteremia (PETERS et al., 2004). Quando causada por bactérias Gram-positivas (GP) ou Gram-negativas (GN), essa condição pode evoluir para sepse (HUTTUNEN; AITTONIEMI, 2011).

A sepse é a principal causa de morte por infecção. Ela ocorre quando o organismo apresenta uma resposta desregulada a uma infecção, levando a uma disfunção orgânica com risco de vida. Se não for identificada e tratada precocemente, pode desencadear choque séptico, falência múltipla dos órgãos e morte. Por isso, é essencial um diagnóstico rápido para que o tratamento adequado seja iniciado o quanto antes (SINGER et al., 2016; EVANS et al., 2021).

As infecções por bactérias Gram-positivas ou Gram-negativas têm origens distintas e ativam respostas imunológicas específicas, o que influencia diretamente seu tratamento (SRISKANDAN; COHEN, 1999). O termo Gram tem origem no nome do pesquisador Christian Gram, responsável por desenvolver a técnica de coloração utilizada na identificação de bactérias. Essa metodologia permite classificar os microrganismos em dois grandes grupos: Gram-positivas ou Gram-negativas. Essa diferenciação está relacionada à composição química das paredes celulares, sendo que as bactérias Gram-negativas apresentam uma estrutura mais complexa do que as Gram-positivas, conforme é apresentado na Figura 1. Essa diferença estrutural interfere diretamente na escolha dos agentes antimicrobianos, pois impacta o mecanismo de ação dos antibacterianos utilizados no tratamento das infecções (ALTERTHUM, 2015).

Essas particularidades estruturais refletem não apenas na interação com o hospedeiro, mas também na forma como esses microrganismos respondem a agentes antimicrobianos. As bactérias Gram-positivas possuem uma parede celular formada por uma camada espessa de peptidoglicano, reforçada por ácidos teicoicos e lipoteicoicos, que contribuem para a rigidez estrutural e ativam a resposta imune do hospedeiro. Já as Gram-negativas apresentam uma camada de peptidoglicano muito mais fina, protegida por uma membrana externa que contém lipopolissacarídeos (LPS), conhecidos como endotoxinas, capazes de desencadear respostas

inflamatórias intensas. Além disso, essa membrana funciona como uma barreira natural contra diversos antibióticos, tornando as infecções por bactérias Gram-negativas geralmente mais difíceis de tratar (WILLEY; SANDMAN; WOOD, 2020).

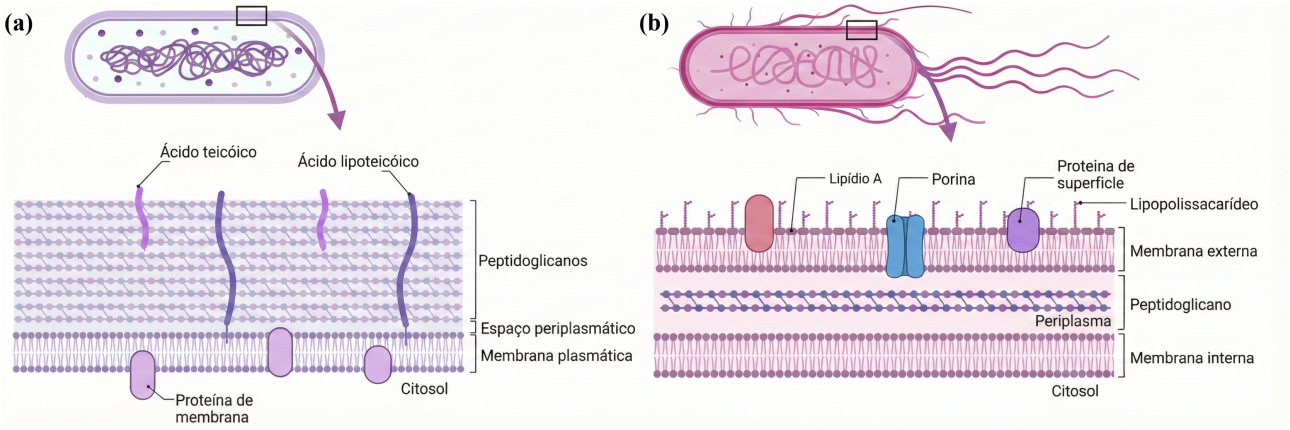

Figura 1 – **(a)** Parede celular de bactéria Gram-positiva; **(b)** Parede celular de bactéria Gram-negativa. Fonte: elaborada pela autora no BioRender.com (2026), adaptado de Willey, Sandman e Wood (2020).

Outros aspectos também distinguem os dois grupos. As bactérias Gram-positivas podem formar endósporos, estruturas altamente resistentes que permitem sua sobrevivência em condições ambientais desfavoráveis, algo que não ocorre nas Gram-negativas. Já as bactérias Gram-negativas apresentam maior diversidade de apêndices de superfície, como fímbrias e pili, que favorecem a adesão a tecidos e a troca de material genético entre células. Essas diferenças ajudam a explicar variações na persistência, transmissão e resistência de cada grupo bacteriano (PRESCOTT; HARLEY; KLEIN, 2002).

No contexto clínico, tais distinções têm consequências importantes. Segundo Tang et al. (2023), a sepse causada por bactérias Gram-negativas apresenta maiores concentrações séricas de fatores inflamatórios e tende a ser mais grave do que a causada por bactérias Gram-positivas. Compreender as particularidades de cada etiologia é fundamental para os profissionais de saúde, pois auxilia na escolha e administração da terapia antimicrobiana empírica apropriada (YANG et al., 2020).

Diante da gravidade dessas infecções e da necessidade de intervenção precoce, o diagnóstico rápido é essencial. O principal e mais confiável método para detectar infecções da corrente sanguínea é a hemocultura (exame laboratorial). No entanto, o processo pode levar de 24 a 72 horas para identificar o patógeno. Como o sucesso do tratamento depende da rapidez do diagnóstico e da administração do antibiótico adequado, diversos esforços têm sido direcionados para aumentar a sensibilidade do método e acelerar a detecção dessas infecções, permitindo um tratamento mais eficaz e reduzindo a morbidade e a mortalidade (PETERS et al., 2004; LOONEN et al., 2014).

Nesse cenário, biomarcadores laboratoriais, como proteína C-reativa (PCR), linfócitos, monócitos, fosfato, plaquetas, creatinina e lactato, têm sido investigados como possíveis preditores na distinção entre infecções causadas por bactérias Gram-positivas ou Gram-negativas (RATZINGER et al., 2015). Paralelamente, cresce o interesse no uso de modelos de aprendizado estatístico no contexto do controle de infecções bacterianas. Esses algoritmos vêm sendo aplicados para modelagem preditiva, avaliação de risco e aprimoramento da precisão diagnóstica (ABU-EL-RUZ et al., 2025).

Apesar dos avanços, a maioria dos modelos preditivos existentes ainda apresenta desempenho moderado ou depende de biomarcadores que não são universalmente acessíveis, como a procalcitonina (PCT), limitando sua aplicação na prática clínica rotineira. Considerando essa lacuna, o presente trabalho busca investigar se a combinação de biomarcadores de baixo custo com métodos de aprendizado estatístico, aliada à avaliação da estabilidade na seleção de variáveis, pode aprimorar a diferenciação precoce entre bacteremias Gram-positivas e Gram-negativas, favorecendo a escolha mais adequada da terapia antimicrobiana inicial.

1.1 Objetivo Geral

Desenvolver e avaliar modelos preditivos baseados em aprendizado estatístico, utilizando biomarcadores laboratoriais de rotina, para a diferenciação precoce entre infecções da corrente sanguínea causadas por bactérias Gram-positivas e Gram-negativas, incorporando critérios de estabilidade na seleção de variáveis.

i. Realizar análise descritiva e exploratória dos biomarcadores laboratoriais, comparando os grupos de pacientes.

ii. Investigar a estabilidade da seleção de variáveis por meio de estratégias de reamostragem, identificando preditores robustos.

iii. Construir modelos de aprendizado estatístico para a classificação da etiologia da bacteremia a partir de diferentes conjuntos de variáveis.

iv. Avaliar e comparar o desempenho discriminatório dos modelos desenvolvidos.

v. Analisar o impacto da parcimônia e da estabilidade dos preditores sobre a capacidade preditiva dos modelos.

1.2 Estrutura do Trabalho

No Capítulo 2, apresenta-se uma revisão de literatura sobre a diferenciação entre bacteremias causadas por bactérias Gram-positivas e Gram-negativas. O Capítulo 3 descreve o banco

de dados utilizado, bem como os procedimentos de pré-processamento e a análise descritiva das variáveis.

No Capítulo 4, avalia-se a estabilidade da seleção de variáveis sob diferentes abordagens e esquemas de reamostragem, com o objetivo de identificar preditores robustos para a discriminação entre os grupos etiológicos. Por fim, o Capítulo 5 apresenta o desenvolvimento e a comparação de modelos preditivos ajustados a partir de distintos conjuntos de variáveis, analisando o impacto da parcimônia e da estabilidade no desempenho discriminatório.

Capítulo 2

Revisão de Literatura

Diversas pesquisas têm sido desenvolvidas com o objetivo de antecipar o diagnóstico de infecções da corrente sanguínea (ICS) por meio de biomarcadores laboratoriais e modelos preditivos baseados em dados rotineiramente disponíveis na prática clínica. A maioria desses estudos busca identificar parâmetros capazes de indicar precocemente a presença de bacteremia e sepse, permitindo intervenções mais rápidas e eficazes no tratamento de pacientes hospitalizados. Podemos destacar Juutilainen et al. (2011), Wang et al. (2013), Ratzinger et al. (2014), Nishikawa et al. (2016), Shim et al. (2019), Steenkiste et al. (2019), Lee et al. (2019), Goh et al. (2021), Su et al. (2021), Hu et al. (2022), Lien et al. (2022), Matono et al. (2022), Schinkel et al. (2022), Uyar et al. (2023), Zhou et al. (2023), Murri et al. (2024), O'Reilly, McGrath e Martin-Loeches (2024), Brouwer, Cunney e Drew (2024), Liu et al. (2025).

Entretanto, uma etapa mais específica e potencialmente decisiva para a escolha da terapia antimicrobiana inicial consiste na diferenciação entre infecções da corrente sanguíne causadas por bactérias Gram-positivas e Gram-negativas. Alguns estudos têm investigado justamente esse aspecto, buscando biomarcadores ou algoritmos capazes de auxiliar na discriminação precoce entre essas etiologias. A seguir, são apresentados alguns dos trabalhos que exploraram essa abordagem.

Estudos iniciais avaliaram a aplicação de algoritmos em parâmetros laboratoriais de rotina. Ratzinger et al. (2015), por exemplo, testaram diferentes técnicas de aprendizado estatístico, alcançando melhor desempenho com o algoritmo K-star (ASC-ROC = 0,675), embora com sensibilidade limitada (0,446), evidenciando a insuficiência de variáveis isoladas para predizer com precisão a etiologia da bacteremia.

Avanços posteriores concentraram-se na avaliação de biomarcadores específicos, especialmente a Procalcitonina (PCT). Lin et al. (2017) mostraram que a Procalcitonina apresentou melhor capacidade discriminativa em comparação com o Lactato e a Proteína C-reativa de alta sensibilidade (hs-CRP), reforçando seu potencial como marcador diagnóstico. De modo semelhante, Thomas-Rüddel et al. (2018) identificaram que níveis de Procalcitonina foram sig-

nificativamente mais elevados em bacteremias Gram-negativas (ASC-ROC = 0,72), ainda que influenciados pelo patógeno e pelo foco infeccioso.

Outros trabalhos exploraram combinações de biomarcadores. Colak et al. (2020) identificaram Procalcitonina, Proteína C-reativa (PCR), Largura de Distribuição dos Eritrócitos (RDW), Largura de Distribuição Plaquetária (PDW), Volume Plaquetário Médio (MPV), Razão Neutrófilo-Linfócito (NLR) e Razão Plaqueta-Linfócito (PLR) como preditores significativos da positividade da hemocultura, com a Procalcitonina mostrando maior acurácia na detecção de Gram-negativas (ASC-ROC = 0,84).

Em linha semelhante, Gao et al. (2021) destacaram que o Volume Plaquetário Médio, Razão Neutrófilo-Linfócito e Procalcitonina foram preditores independentes de bacteremia Gram-negativas, resultando em um modelo com ASC-ROC = 0,79, sensibilidade de 0,75 e especificidade de 0,71. Além disso, Tang et al. (2020) demonstraram que parâmetros hematológicos simples, como Razão Neutrófilo-Linfócito e Percentual de Plaquetas Grandes (P-LCR), quando combinados, podem alcançar alto poder discriminativo (ASC-ROC até 0,943 para *Enterobacter aerogenes*), reforçando o potencial de biomarcadores de baixo custo.

A Procalcitonina aparece de forma recorrente nesses estudos como marcador promissor, inclusive em contextos específicos como pacientes cirúrgicos sépticos (NEŠKOVIĆ et al., 2023) e em análises multigrupos que incluem fungos (BASSETTI et al., 2020). Apesar de sua utilidade, o exame possui limitações práticas, como alto custo, necessidade de equipamentos especializados e a exigência de dosagens seriadas, fatores que restringem sua aplicação em hospitais com restrições financeiras e reforçam a necessidade de alternativas viáveis (LOONEN et al., 2014; LIEN et al., 2022).

Paralelamente, cresce o interesse no uso de modelos de aprendizado estatístico. Mahmoud et al. (2021) desenvolveram modelos preditivos para hemoculturas positivas e negativas utilizando variáveis clínicas e laboratoriais de rotina, alcançando desempenho global limitado (ASC-ROC $<$ 0,54), apesar de apresentar maior especificidade que sensibilidade. Já Zhang et al. (2023) aplicaram diferentes algoritmos, com destaque para florestas aleatórias, que obtiveram ASC-ROC = 0,768, além de propor uma árvore de decisão simplificada com cinco variáveis (Contagem de Glóbulos Brancos, Porcentagem de Basófilos, Fosfatase Alcalina, Lactato e Bilirrubina Total) acessíveis à prática hospitalar com ASC-ROC = 0,679.

Mais recentemente, Wang et al. (2025) investigaram a identificação precoce de bacteremia por meio de modelos preditivos baseados em múltiplos biomarcadores clínicos e laboratoriais de rotina, observando melhor desempenho quando os preditores são avaliados de forma integrada. Esses resultados reforçam a relevância de abordagens multivariáveis e evidenciam a importância de estratégias metodológicas capazes de identificar conjuntos de variáveis informativas e reprodutíveis em dados clínicos complexos.

Por sua vez, Dambroso-Altafini et al. (2022) desenvolveram um modelo de regressão logística partindo de 68 biomarcadores laboratoriais, dos quais apenas quatro permaneceram no modelo final (Plaquetas, Creatinina, Hemoglobina Corpuscular Média e Eritrócitos), alcançando ASC-ROC = 0,69. Este estudo é particularmente relevante para a presente dissertação, pois utilizou o mesmo banco de dados e reforça a necessidade de explorar algoritmos mais robustos para aprimorar a capacidade discriminatória.

Embora diversos estudos tenham obtido resultados promissores na diferenciação entre bacteremias Gram-positivas e Gram-negativas, a maior parte das investigações concentra-se prioritariamente no desempenho preditivo final dos modelos, com menor ênfase na avaliação sistemática da estabilidade da seleção de variáveis. A literatura metodológica destaca que a estabilidade constitui um aspecto relevante na modelagem orientada por dados, especialmente em cenários com múltiplos preditores e potencial colinearidade (HEINZE; WALLISCH; DUNKLER, 2018; NOGUEIRA; SECHIDIS; BROWN, 2018). Em contextos clínicos com biomarcadores correlacionados e tamanho amostral limitado, pequenas variações na amostra podem influenciar o conjunto de variáveis selecionadas (BREIMAN, 1996b; RILEY; COLLINS, 2023).

Além disso, modelos avaliados apenas pelo desempenho aparente podem apresentar estimativas otimistas quando não submetidos a estratégias adequadas de validação interna (HARRELL, 2015; STEYERBERG, 2019). Nesse sentido, a integração entre técnicas de reamostragem e métodos de regularização tem sido apontada como uma abordagem promissora para reduzir a variabilidade da seleção e favorecer a identificação de um núcleo preditivo mais consistente (TIBSHIRANI, 1996; MEINSHAUSEN; BÜHLMANN, 2010; SAUERBREI et al., 2015).

De modo geral, a literatura ainda carece de investigações que integrem explicitamente desempenho preditivo, estabilidade da seleção e parcimônia na construção de modelos clínicos. Nesse contexto, parâmetros hematológicos e bioquímicos de baixo custo e ampla disponibilidade apresentam-se como alternativas promissoras. Para contribuir com esse avanço, o presente estudo investiga biomarcadores laboratoriais de rotina em conjunto com algoritmos de aprendizado estatístico, incorporando a estabilidade e a parcimônia como princípios metodológicos centrais. Dessa forma, avalia-se como esses critérios influenciam a construção dos modelos preditivos, com o propósito de aprimorar a diferenciação precoce entre bacteremias Gram-positivas e Gram-negativas e ampliar sua aplicabilidade clínica.

Capítulo 3

Dados, Análise Exploratória e Pré-processamento

Este capítulo descreve a base de dados utilizada no estudo e as etapas de análise exploratória e pré-processamento realizadas antes da análise de estabilidade e da modelagem preditiva. São apresentados a origem e as características dos dados, a análise descritiva dos biomarcadores e os critérios adotados para exclusão de dados com alta proporção de valores ausentes e controle de multicolinearidade, que resultaram na definição da base final. Todas as etapas de tratamento e análise de dados foram realizadas utilizando o *software* **R** (R Core Team, 2024), com pacotes específicos para manipulação, pré-processamento e visualização de dados.

3.1 Dados

Os dados utilizados nesta dissertação foram fornecidos por Dambroso-Altafini et al. (2022) e são provenientes do Hospital Universitário de Maringá (HUM). No estudo original, 455 pacientes atendidos entre 2013 e 2018 foram utilizados para a construção do modelo preditivo, enquanto os registros de 2019 foram reservados para validação.

Neste trabalho, todos os registros disponíveis entre 2013 e 2019 foram consolidados em uma única base de dados. Após essa unificação, a amostra final foi composta por 523 pacientes com infecção da corrente sanguínea (ICS), dos quais 246 foram causadas por bactérias Gram-positivas (GP) e 277 por Gram-negativas (GN).

A decisão de agrupar os dados de todos os anos teve como objetivo maximizar o tamanho amostral e ampliar a disponibilidade de informações para a modelagem. Posteriormente, a base completa foi dividida aleatoriamente em conjuntos de treinamento e validação interna para aplicação dos algoritmos de aprendizado estatístico. A descrição detalhada das variáveis utilizadas encontra-se no Apêndice B.

3.2 Análise Exploratória dos Dados

Inicialmente, foi realizada uma análise exploratória considerando todos os biomarcadores disponíveis na base original. Foram examinadas distribuições, padrões e capacidade discriminatória dos biomarcadores. Para as variáveis contínuas, foram geradas estimativas de densidade *Kernel*, que fornecem uma representação suave das distribuições. Além disso, foram calculadas medidas resumo (média, desvio padrão e percentis) e métricas de discriminação individual, como Área Sob a Curva ROC (ASC-ROC), obtida por meio da função `roc` do pacote `pROC` (ROBIN et al., 2011), e distância L^2.

A ASC-ROC varia de 0,5 (ausência de discriminação) a 1,0 (discriminação perfeita) e pode ser interpretada como a probabilidade de que um indivíduo do grupo positivo apresente valor superior ao de um indivíduo do grupo negativo (HANLEY; MCNEIL, 1982). Valores próximos de 0,5 indicam baixa capacidade discriminatória, enquanto valores mais elevados refletem melhor distinção entre os grupos (STEYERBERG et al., 2010).

A distância L^2 quantifica a separação entre as distribuições estimadas dos grupos. Valores menores indicam maior sobreposição e, portanto, menor poder discriminatório, enquanto valores mais elevados sugerem maior separação entre os grupos (THENG; BHOYAR, 2024). Diferentemente da ASC-ROC, não existem pontos de corte universalmente estabelecidos para classificar valores de L^2 como baixos ou altos, a sua interpretação deve ser feita de forma comparativa, considerando a escala e o conjunto de variáveis analisado.

No presente estudo, algumas variáveis apresentaram melhor desempenho discriminatório individual. A Creatinina (ASC-ROC = 0,636; L^2 = 0,00520; Tabela 14) e o Lactato (ASC-ROC = 0,614; L^2 = 0,00513; Tabela 18), por exemplo, exibiram maior distinção entre as distribuições dos grupos, sugerindo potencial discriminatório individual. Entretanto, a maioria dos biomarcadores apresentou valores de ASC-ROC próximos de 0,5 e L^2 na ordem de 0,001 a 0,002, indicando maior sobreposição entre as distribuições. Como exemplo, a Hemoglobina Corpuscular Média (HCM) apresentou ASC-ROC de 0,563 e L^2 = 0,00141 (Tabela 4), evidenciando baixa discriminação isolada entre os grupos Gram-positivos e Gram-negativos.

Esses resultados indicam que, embora algumas variáveis apresentem alguma capacidade discriminatória quando analisadas isoladamente, a distinção entre os grupos com base em um único biomarcador é limitada. Achados semelhantes foram observados para a maioria dos biomarcadores, reforçando a necessidade de abordagem multivariável capaz de integrar informações de múltiplos preditores para melhorar o desempenho diagnóstico. As demais variáveis e seus respectivos valores de ASC-ROC e L^2 encontram-se detalhados no Apêndice A. A partir dessa análise, foram realizadas as etapas de pré-processamento descritas a seguir.

3.3 Pré-processamento dos Dados

Após a etapa inicial de análise exploratória descritiva, foram aplicados procedimentos de pré-processamento com o objetivo de reduzir a esparsidade da matriz de dados e garantir a robustez dos modelos preditivos. Para essa etapa, considerou-se conjuntamente a presença de valores ausentes e valores iguais a zero, uma vez que ambos indicam ausência ou raridade da informação observada para determinados biomarcadores. Assim, adotaram-se os seguintes critérios, nesta ordem:

1. **Remoção de indivíduos:** pacientes com 80% ou mais valores ausentes ou nulos nos biomarcadores foram removidos, garantindo registros suficientes para modelagem confiável.
2. **Exclusão de variáveis:** biomarcadores com pelo menos 60% de valores ausentes ou nulos foram descartados, mantendo apenas variáveis com representatividade adequada.
3. **Tratamento de multicolinearidade:** análise de correlação de Pearson entre as variáveis remanescentes foi realizada; variáveis com correlação absoluta superior a 0,85 foram removidas utilizando a função `findCorrelation` do pacote `caret` (KUHN, 2008).

A escolha do limiar de 60% de valores ausentes para exclusão de variáveis representou um compromisso entre manter um número adequado de pacientes e evitar o uso extensivo de marcadores com baixa disponibilidade de dados. Com base nesse critério, exames de coagulação como Tempo de Atividade da Protrombina (TAP) e TAP RNI (Razão Normalizada Internacional), que apresentaram proporções de dados faltantes em torno de 67%, foram excluídos da base final, apesar de exibirem discriminação individual moderada (ASC-ROC em torno de 0,67; Tabelas 28 e 29). Assim, optou-se por concentrar a modelagem em biomarcadores laboratoriais de rotina com maior quantidade de informação, deixando a investigação específica desses exames como linha de pesquisa futura.

Esses critérios foram adotados com base em recomendações da literatura metodológica sobre modelagem preditiva, que enfatiza a necessidade de controle da alta proporção de dados ausentes e da multicolinearidade entre preditores. Essas medidas visam aumentar a estabilidade dos modelos, reduzir distorções nas estimativas e evitar problemas decorrentes da alta correlação entre preditores, contribuindo para melhor desempenho e maior capacidade de generalização (HARRELL, 2015; DORMANN et al., 2013; HEINZE; WALLISCH; DUNKLER, 2018). Após essas etapas, a base final consistiu em 500 pacientes e 38 variáveis preditoras. Entre os indivíduos incluídos, 237 (47,4%) apresentaram bacteremia causada por bactérias Gram-positivas (GP) e 263 (52,6%) por Gram-negativas (GN).

Capítulo 4

Análise de Estabilidade na Seleção de Variáveis

4.1 Introdução

A seleção de variáveis é uma etapa essencial na modelagem estatística, voltada à construção de modelos parcimoniosos e interpretáveis, sem comprometer a capacidade preditiva. Como destacou Box (1979), nenhum modelo descreve perfeitamente a realidade, mas alguns são úteis para compreender fenômenos e apoiar decisões.

A escolha do subconjunto de variáveis envolve um equilíbrio delicado: incluir muitas aumenta a variância das estimativas, enquanto excluir em excesso introduz viés. Além de melhorar a interpretação e reduzir custos de coleta, a seleção adequada ajuda a evitar o sobreajuste (MILLER, 2002). Trata-se, portanto, de um desafio estatístico relevante, que exige soluções capazes de equilibrar simplicidade, desempenho e estabilidade (GEORGE, 2000).

Os métodos sequenciais tornaram-se amplamente utilizados devido à sua simplicidade e disponibilidade em *softwares* estatísticos. No entanto, sua popularidade contrasta com as críticas recorrentes na literatura. Estudos apontam que pequenas variações nos dados podem modificar substancialmente o conjunto final de variáveis, comprometendo a estabilidade e a generalização dos modelos (HOCKING, 1976; DERKSEN; KESELMAN, 1992; BREIMAN, 1996b).

Uma alternativa para investigar essa instabilidade é o uso de reamostragem. O *bootstrap*, proposto por Efron (1979b), passou a ser utilizado não apenas para estimar incerteza, mas também para avaliar consistência na seleção de variáveis. Aplicações em estudos clínicos destacaram a estabilidade como critério essencial de qualidade (SAUERBREI; ROYSTON, 1999; SAUERBREI et al., 2015).

Posteriormente, abordagens baseadas em regularização ampliaram o escopo da seleção. O LASSO (*Least Absolute Shrinkage and Selection Operator*), proposto por Tibshirani (1996)

introduziu penalização para simplificar modelos e reduzir instabilidade, e a *Stability Selection* (MEINSHAUSEN; BÜHLMANN, 2010) combinou penalização e reamostragem para reforçar a consistência dos preditores selecionados.

Neste capítulo, avalia-se a estabilidade da seleção de variáveis em dois cenários: um baseado em métodos sequenciais (*stepwise*) e outro em regressão penalizada LASSO. Em ambos os casos, a estabilidade é investigada por meio de reamostragem, com o objetivo de quantificar a consistência da seleção e identificar preditores robustos para a diferenciação entre bacteremias Gram-positivas e Gram-negativas, contribuindo para a construção de modelos mais confiáveis.

4.2 Fundamentação Teórica

A seleção de variáveis é uma etapa central no desenvolvimento de modelos preditivos, pois influencia diretamente o desempenho, a interpretabilidade e a aplicabilidade dos modelos (BOX, 1979; ZOU; HASTIE, 2005; HEINZE; WALLISCH; DUNKLER, 2018). Em estudos baseados em biomarcadores laboratoriais de rotina, o conjunto inicial de preditores costuma ser amplo e frequentemente inclui variáveis correlacionadas ou com contribuição preditiva limitada.

Nesse contexto, o processo de seleção exige equilibrar inclusão e exclusão de preditores, evitando tanto sobreajuste quanto omissão de variáveis relevantes (MURTAUGH, 1998; SAUERBREI et al., 2015). Em cenários com muitos preditores, a escolha adequada é ainda mais crítica (YIN; LI; ZHANG, 2017). Além disso, modelos úteis devem ser estáveis, ou seja, apresentar resultados consistentes frente a pequenas variações nos dados (YU, 2013).

4.2.1 Conceito de Estabilidade

A estabilidade refere-se à capacidade de um método estatístico produzir resultados consistentes quando pequenas mudanças são feitas nos dados. Em termos práticos, um procedimento é estável quando os modelos, coeficientes e variáveis selecionadas permanecem semelhantes sob perturbações leves na amostra.

Segundo Breiman (1996b), algoritmos estatísticos diferem quanto à sensibilidade a ruídos e flutuações amostrais. Procedimentos instáveis, como seleção de subconjuntos, podem resultar em modelos muito diferentes a partir de amostras quase idênticas, levando à perda de desempenho preditivo e maior variância. De forma complementar, Yu (2013) argumenta que estabilidade é requisito fundamental para inferência confiável, pois modelos estáveis tendem a capturar padrões reais, enquanto métodos instáveis evidenciam variações aleatórias.

A estabilidade pode ser analisada sob diferentes perspectivas. Na estimação, ela corresponde à consistência dos coeficientes ajustados sob pequenas alterações nos dados. Yu (2013) e Lim e Yu (2016) propuseram métricas específicas, como o critério ESCV, que combina validação

cruzada e estabilidade para favorecer modelos esparsos e menos sensíveis ao ruído. Na predição, estabilidade refere-se à robustez das probabilidades ou medidas de risco geradas por um modelo quando aplicado a amostras levemente diferentes, aspecto relevante na prática clínica (RILEY; COLLINS, 2023).

Por fim, a estabilidade da seleção de variáveis, foco deste estudo, diz respeito à frequência com que os mesmos preditores são escolhidos em diferentes reamostragens. Trabalhos como Altman e Andersen (1989) e Sauerbrei e Schumacher (1992) introduziram o uso do *bootstrap* para investigar essa estabilidade, mostrando que variáveis realmente informativas tendem a ser selecionadas repetidamente, enquanto variáveis instáveis refletem flutuações amostrais. Nesse contexto, estabilidade não é apenas uma propriedade desejável, mas um critério de qualidade do modelo, essencial para garantir reprodutibilidade e interpretação confiável (SAUERBREI et al., 2015).

4.2.2 Instabilidade dos Métodos Sequenciais

Procedimentos clássicos de seleção de variáveis amplamente conhecidos incluem a eliminação *backward* (regressiva), seleção *forward* (progressiva) e *stepwise* (bidirecional). Na eliminação *backward*, parte-se do modelo completo e variáveis são removidas progressivamente; na seleção *forward*, inicia-se do modelo nulo e variáveis são adicionadas sucessivamente; já o *stepwise* combina ambas as abordagens, permitindo inclusões e exclusões ao longo do processo (DERKSEN; KESELMAN, 1992). Todas essas estratégias são conduzidas com base em um critério de seleção, que pode ser definido por níveis de significância estatística (p-valores) ou por medidas de informação, como o Critério de Informação de Akaike (AIC) (AKAIKE, 1974) ou o Critério Bayesiano (BIC) (SCHWARZ, 1978), visando equilibrar ajuste e parcimônia do modelo.

Esses métodos são amplamente utilizados pela simplicidade e pela implementação direta em *softwares* estatísticos (FLACK; CHANG, 1987; DESBOULETS, 2018). Apesar da popularidade, há consenso na literatura de que esses procedimentos apresentam limitações importantes, especialmente quanto à consistência e à qualidade inferencial dos modelos resultantes.

Um ponto central é que esses métodos não garantem a obtenção do melhor subconjunto de variáveis entre todas as combinações possíveis. Beale (1970) e Mantel (1970) demonstraram que o *stepwise* pode convergir para soluções locais, já que a inclusão ou exclusão sucessiva de variáveis não assegura uma solução globalmente ótima. De modo semelhante, Hocking (1976) observou que diferentes ordens de entrada podem gerar modelos distintos, refletindo a natureza heurística do algoritmo. Além disso, a inclusão de uma variável por vez pode impedir a detecção de combinações relevantes de preditores (MANTEL, 1970).

Outro conjunto de críticas refere-se à inferência estatística. Copas (1983) apontou que, ao

reutilizar os mesmos dados para selecionar e estimar o modelo, ocorre superestimação do ajuste e subestimação dos erros. Miller (2002) mostrou que o viés de seleção afasta os coeficientes de zero e torna os valores de p pouco confiáveis. Trabalhos posteriores como Burnham e Anderson (2002), Heinze e Dunkler (2017), reforçam que coeficientes e testes de hipótese após seleção tendem a ser enviesados, pois a incerteza sobre o processo de escolha do modelo é ignorada, levando a interpretações equivocadas.

Há também limitações metodológicas adicionais. Hocking (1976) destacou que a ordem de entrada ou saída das variáveis reflete apenas a lógica do algoritmo, e não sua relevância real. A escolha dos níveis de significância para inclusão ou exclusão é subjetiva e influencia diretamente o modelo final, podendo gerar resultados diferentes com pequenas alterações nesses limiares. A multiplicidade de testes, já mencionada por Tukey (1977), aumenta a probabilidade de inclusão de variáveis irrelevantes. Além disso, métodos como *forward* e *backward* não reavaliam variáveis previamente excluídas, o que pode levar à omissão de preditores importantes (DERKSEN; KESELMAN, 1992; DESBOULETS, 2018). Esses fatores contribuem para modelos excessivamente adaptados (sobreajustados) aos dados da amostra e com baixa capacidade de generalização (BURNHAM; ANDERSON, 2002).

Entre todas as limitações destacadas, a instabilidade da seleção é reconhecida como uma das mais críticas. Breiman (1996b) classificou a seleção de subconjuntos como notoriamente instável, resultando em perda de desempenho preditivo e aumento da variabilidade do erro. Evidências empíricas reforçam essa preocupação. Altman e Andersen (1989) mostraram, por meio de reamostragem *bootstrap* em modelos de Cox, que pequenas mudanças nos dados podem gerar conjuntos de variáveis completamente distintos. Estudos de simulação, como Flack e Chang (1987) e Derksen e Keselman (1992), revelaram que uma proporção substancial das variáveis selecionadas corresponde a ruído puro, chegando a ultrapassar 80% em determinados cenários. Trabalhos posteriores confirmaram essa fragilidade e destacaram a importância de avaliar a estabilidade das variáveis incluídas (SAUERBREI; SCHUMACHER, 1992; STEYERBERG et al., 2001).

Diante dessas limitações, recomenda-se cautela no uso exclusivo de métodos sequenciais de seleção de variáveis. A literatura destaca que, além do desempenho preditivo, é fundamental considerar também a estabilidade das variáveis selecionadas ao longo de diferentes amostras (MILLER, 2002; SAUERBREI et al., 2015). Nesse contexto, abordagens baseadas em regularização, como o LASSO, e métodos como a *Stability Selection* têm sido propostas como alternativas mais robustas.

Além disso, técnicas de reamostragem passaram a desempenhar papel importante na avaliação da consistência do processo de seleção. Entre elas, o *bootstrap* permite investigar com que frequência cada variável é selecionada em diferentes amostras, fornecendo uma medida prática de estabilidade e auxiliando na identificação de preditores mais confiáveis.

4.2.3 Avaliação da Estabilidade por Reamostragem

A reamostragem tornou-se uma ferramenta essencial para avaliar a estabilidade da seleção de variáveis, especialmente porque muitas técnicas automáticas tendem a escolher preditores diferentes quando a amostra muda ligeiramente. Essa sensibilidade dificulta identificar quais variáveis são realmente relevantes e pode comprometer a validade do modelo em novos dados. Ao reproduzir o processo de seleção em várias amostras derivadas dos dados originais, é possível medir o quanto os resultados dependem da amostra específica e não dos padrões reais do fenômeno.

O método mais utilizado para avaliar estabilidade é o *bootstrap*, introduzido por Efron (1979a). A ideia consiste em reamostrar a base original com reposição, obtendo múltiplas amostras do mesmo tamanho. Em cada reamostragem, o modelo é ajustado novamente, permitindo observar a sensibilidade do processo de seleção a pequenas mudanças nos dados. Além disso, o *bootstrap* funciona como validação interna, corrigindo o otimismo do ajuste quando o modelo é avaliado na própria amostra (EFRON; GONG, 1983).

O uso do *bootstrap* para investigar variabilidade na seleção foi explorado por Gong (1982) e Chen e George (1985), que propuseram quantificar a frequência de seleção de cada variável. Esse conceito, conhecido como *Bootstrap Inclusion Frequency* (BIF), foi consolidado em aplicações práticas como as de Altman e Andersen (1989) e Sauerbrei e Schumacher (1992), que mostraram que pequenas alterações nos dados podem alterar de forma expressiva o conjunto de preditores escolhidos.

Outra abordagem para avaliar estabilidade é o *subsampling*, no qual se obtêm amostras menores do conjunto original, sem reposição. Diferentemente do *bootstrap*, que procura reproduzir a distribuição empírica dos dados, o *subsampling* introduz perturbações mais intensas ao trabalhar com subconjuntos reduzidos. Seus fundamentos foram desenvolvidos em trabalhos clássicos de Hartigan (1969), Hartigan (1975) e posteriormente formalizados por Politis e Romano (1994) e Politis, Romano e Wolf (1999), que demonstraram a validade assintótica da abordagem sob condições gerais.

A escolha do tamanho da subamostra é um aspecto importante, pois afeta o grau de perturbação aplicado ao conjunto de dados. Denotando por n o tamanho da amostra original e por m o tamanho da subamostra, quando m cresce com n, mas em velocidade mais lenta, o *subsampling* apresenta boas propriedades assintóticas e pode superar o *bootstrap* em situações de alta variabilidade ou dependência entre observações. Em contextos de aprendizado estatístico, a subamostragem também atua como uma forma de regularização, reduzindo a influência de padrões específicos da amostra e aumentando a estabilidade da seleção.

A frequência de inclusão resultante das reamostragens tornou-se uma métrica amplamente utilizada para quantificar estabilidade. Trabalhos como Sauerbrei, Boulesteix e Binder

(2011) e Sauerbrei et al. (2015) mostram que variáveis verdadeiramente informativas tendem a ser selecionadas repetidamente, enquanto preditores instáveis aparecem com baixa frequência. Essa abordagem é amplamente utilizada em estudos clínicos e epidemiológicos para identificar fatores prognósticos consistentes (STEEN et al., 2002; AUSTIN; TU, 2004; BEYENE et al., 2009; PAIVA et al., 2021; HYDE et al., 2021; OTTARSDOTTIR et al., 2022; HUMMEL et al., 2025).

Estudos comparativos também avaliaram o desempenho de diferentes estratégias de reamostragem no contexto de seleção de variáveis. Bin et al. (2016) avaliaram o desempenho do *Bootstrap* e *Subsampling* em diferentes cenários e mostraram que o *Subsampling* tende a produzir seleções mais conservadoras e estáveis, enquanto o *Bootstrap* pode incluir variáveis menos robustas. Esses resultados reforçam a utilidade de comparar diferentes esquemas de reamostragem, como será feito neste trabalho.

Em síntese, técnicas de reamostragem oferecem uma base sólida para avaliar a confiabilidade da seleção de variáveis, permitindo distinguir preditores consistentemente relevantes de efeitos decorrentes do acaso ou de particularidades da amostra.

4.2.4 Regularização e Estabilidade

A regularização surgiu como alternativa aos métodos sequenciais, pois permite controlar a complexidade do modelo de forma contínua, sem decisões abruptas de inclusão ou exclusão de variáveis. Em vez disso, ela reduz gradualmente o tamanho dos coeficientes, favorecendo modelos mais estáveis. Essa ideia foi introduzida por Hoerl e Kennard (1970) com a regressão *Ridge*, que mostrou que introduzir um viés controlado nas estimativas pode reduzir a variância e melhorar o desempenho preditivo, especialmente quando há tamanho amostral reduzido ou forte correlação entre preditores.

O LASSO, proposto por Tibshirani (1996), expandiu esse conceito ao permitir, além do encolhimento dos coeficientes, a eliminação automática de variáveis. Ao aplicar uma penalização do tipo ℓ_1, o LASSO força alguns coeficientes a zero, combinando ajuste e seleção de variáveis em um único procedimento. As suas estimativas são obtidas como a solução do seguinte problema de otimização:

$$(\hat{\beta}_0, \hat{\beta}) = \arg\min_{\beta_0,\beta} \left\{ \frac{1}{2n} \sum_{i=1}^{n} (y_i - \beta_0 - x_i^\top \beta)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\},$$

em que y_i representa a resposta observada, β_0 é o intercepto (que não sofre penalização), x_i é o vetor de preditores, n é o número de observações, p é o número de variáveis e $\lambda \geq 0$ o parâmetro de regularização que controla a intensidade da penalização. Valores maiores de λ aumentam o grau de encolhimento (*shrinkage*) e produzem modelos mais parcimoniosos, enquanto valores

menores tendem a incluir mais variáveis.

Em situações em que a resposta é binária, o LASSO é aplicado à regressão logística. O modelo logístico assume a forma:

$$log\frac{P(Y=1|X)}{P(Y=0|X)} = \beta_0 + {x_i}^\top \beta,$$

e o LASSO estima os coeficientes minimizando a log-verossimilhança penalizada:

$$\min_{\beta_0,\beta}\left\{-\sum_{i=1}^{n}\left[y_i(\beta_0 + x_i^T\beta) - \log\left(1+e^{\beta_0+x_i^T\beta}\right)\right] + \lambda\sum_{j=1}^{p}|\beta_j|\right\}.$$

Nesse contexto, o primeiro termo mede o ajuste do modelo (via log-verossimilhança), enquanto o segundo impõe a penalização ℓ_1, responsável por reduzir ou anular coeficientes. Essa combinação produz modelos mais estáveis, esparsos e interpretáveis, reduzindo o risco de sobreajuste, especialmente em conjuntos de dados com muitos preditores.

A escolha adequada de λ é fundamental para equilibrar ajuste e parcimônia. Esse parâmetro é normalmente determinado por validação cruzada, que estima o erro de predição médio em uma grade de valores de λ. Dois valores de referência são amplamente utilizados:

- λ_{min}: valor que minimiza o erro de predição médio;
- λ_{1se}: valor que produz o modelo mais parcimonioso cujo erro está dentro de um desvio padrão do mínimo.

O uso conjunto desses dois parâmetros permite comparar modelos que priorizam a acurácia (λ_{min}) e a simplicidade (λ_{1se}), oferecendo uma avaliação mais completa da estabilidade e da relevância dos preditores.

A principal vantagem da penalização do tipo ℓ_1, utilizada no LASSO, é a sua capacidade de reduzir alguns coeficientes exatamente a zero, realizando simultaneamente ajuste e seleção automática de variáveis (TIBSHIRANI, 1996). Em contraste, a penalização do tipo ℓ_2, característica da regressão Ridge (HOERL; KENNARD, 1970), atua apenas no encolhimento dos coeficientes, aproximando-os de zero para reduzir a variância das estimativas e mitigar problemas de multicolinearidade, mas sem eliminá-los completamente, ou seja, ela não realiza seleção automática de preditores.

Métodos relacionados, como o *Elastic Net* (ZOU; HASTIE, 2005), combinam penalizações ℓ_1 e ℓ_2, sendo particularmente úteis em situações com preditores fortemente correlacionados. No entanto, como o objetivo desta etapa é identificar um subconjunto reduzido e parcimonioso de biomarcadores com capacidade preditiva, neste trabalho optou-se por utilizar exclusivamente a penalização ℓ_1 por meio do método LASSO.

Embora o LASSO apresente maior estabilidade que métodos como o *stepwise*, ele ainda pode ser sensível a pequenas variações na amostra, especialmente quando há poucos dados ou alta colinearidade entre preditores. Por isso, a avaliação empírica da estabilidade continua sendo relevante mesmo em modelos penalizados. Estudos como Altman e Andersen (1989) e Sauerbrei e Schumacher (1992) demonstram que técnicas de reamostragem, como o *bootstrap*, permitem identificar variáveis selecionadas de forma consistente ao longo de múltiplas réplicas dos dados.

Uma evolução importante nesse contexto é a *Stability Selection*, proposta por Meinshausen e Bühlmann (2010). Esse método combina regularização com subamostragem repetida para calcular a probabilidade de seleção de cada variável, oferecendo controle explícito sobre falsos positivos e aumentando a confiabilidade dos resultados, especialmente em cenários de alta dimensionalidade.

4.2.5 Parcimônia e Interpretabilidade Clínica

A parcimônia é um princípio central na construção de modelos estatísticos e está diretamente relacionada à estabilidade e à interpretabilidade dos resultados. Modelos excessivamente complexos tendem a apresentar maior variabilidade de desempenho entre amostras e maior risco de sobreajuste, além de dificultarem a interpretação clínica.

Por outro lado, modelos excessivamente simples podem comprometer a capacidade discriminatória. Estudos metodológicos indicam que, em muitos contextos clínicos, é possível obter desempenho preditivo semelhante utilizando um número reduzido de preditores, desde que esses sejam selecionados de forma criteriosa e avaliados quanto à sua estabilidade (RILEY; COLLINS, 2023).

Em aplicações baseadas em biomarcadores laboratoriais, modelos mais simples apresentam ainda vantagens adicionais, pois facilitam a implementação em sistemas de apoio à decisão clínica e a interpretação dos fatores associados ao desfecho de interesse (CHOWDHURY; TURIN, 2020; DIAZ-RAMIREZ et al., 2021; MATHARAARACHCHI; DOMARATZKI; MUTHUKUMARANA, 2022). Assim, a busca por modelos parcimoniosos deve equilibrar desempenho preditivo, estabilidade da seleção de variáveis e interpretabilidade clínica (HARRELL, 2015; KAMKAR et al., 2015).

Nesse contexto, a integração entre regularização, reamostragem e análise de estabilidade constitui uma estratégia metodológica importante para identificar preditores consistentes e construir modelos robustos. No presente trabalho, esses princípios são utilizados para definir os conjuntos de variáveis posteriormente empregados no treinamento e na avaliação dos algoritmos de aprendizado supervisionado apresentados no Capítulo 5.

4.3 Metodologia

Esta seção descreve os procedimentos utilizados para avaliar a estabilidade na seleção de variáveis em modelos de regressão logística. Os dados estão descritos na Seção 3.1 e o pré-processamento na Seção 3.3. As análises foram conduzidas combinando três estratégias de reamostragem: *Bootstrap(n)*, *Bootstrap(m)* e *Subsampling(m)*; com dois métodos de seleção de variáveis: o *stepwise* e a regressão penalizada LASSO. A fundamentação teórica detalha os princípios estatísticos e as formulações matemáticas de cada método.

4.3.1 Reamostragem e Critérios de Estabilidade

A estabilidade foi avaliada repetindo o processo de seleção de variáveis em múltiplas reamostragens dos dados originais, conforme proposto por Bin et al. (2016). Foram utilizadas as seguintes estratégias:

- **Bootstrap(n)**: reamostragem com reposição, B vezes, com o mesmo tamanho da amostra original (n);
- **Bootstrap(m)**: reamostragem com reposição, B vezes, com tamanho reduzido ($m < n$);
- **Subsampling(m)**: reamostragem sem reposição, B vezes, também com tamanho $m < n$.

O tamanho da subamostra foi definido como $m = \lfloor 0{,}632 \cdot n \rfloor$, ou seja, o valor inteiro obtido pelo arredondamento para baixo de $0{,}632 \cdot n$, que corresponde ao número médio de observações únicas em uma amostra *bootstrap*. O número de replicações foi fixado em $B = 5000$.

A estabilidade de cada variável foi quantificada pela frequência de seleção, ou seja, a proporção de vezes em que a variável foi incluída entre os B modelos ajustados. Valores próximos de 1 indicam alta consistência e valores baixos indicam baixa relevância ou instabilidade amostral.

Embora não exista um limiar universal para definir variáveis estáveis, a literatura sobre estabilidade da seleção baseada em frequências de inclusão por reamostragem sugere que o ponto de corte deve situar-se entre 50% e 100%, a fim de reduzir a seleção de variáveis associadas apenas ao ruído amostral (MEINSHAUSEN; BÜHLMANN, 2010). Em aplicações práticas, valores entre 60% e 90% são frequentemente utilizados para diferenciar preditores consistentes daqueles selecionados por flutuações da amostra (SAUERBREI et al., 2015; NOGUEIRA; SECHIDIS; BROWN, 2018).

Com base nessas recomendações, adotou-se neste estudo o valor mínimo de 65% de frequência de inclusão como critério operacional para identificar variáveis mais estáveis. Esse limiar foi utilizado apenas como medida descritiva da estabilidade ao longo das reamostragens.

4.3.2 Seleção de Variáveis via Métodos Clássicos (*Stepwise*)

Os procedimentos de seleção clássicos (*backward*, *forward* e *stepwise*) foram aplicados conforme detalhado na Seção 4.2.2, utilizando como critério de decisão quatro níveis de significância (α): 0,001, 0,05, 0,10 e 0,157. O valor de 0,157 foi incluído por corresponder ao critério de decisão implícito no AIC para inclusão de variáveis em modelos aninhados, reproduzindo o equilíbrio entre qualidade de ajuste e complexidade do modelo, com foco na capacidade preditiva (HARRELL, 2015; HEINZE; WALLISCH; DUNKLER, 2018).

A cada reamostragem, um modelo de regressão logística foi ajustado e as variáveis selecionadas foram registradas, conforme os algoritmos descritos a seguir.

Algorithm 1 Reamostragem com seleção *stepwise*

Input: Dados D; número de replicações B; procedimento de reamostragem (*Bootstrap(n)*, *Bootstrap(m)* ou *Subsampling(m)*); critério de inclusão/exclusão α para seleção *stepwise*;
for $b = 1$ to B **do**
    Reamostrar D_b segundo o procedimento escolhido;
    Ajustar modelo de regressão logística via seleção *stepwise* (α);
    Registrar variáveis selecionadas e atualizar frequências;
end for
Output: Frequência de inclusão das variáveis ao longo das B reamostragens.

4.3.3 Seleção de Variáveis via Regularização LASSO

Na segunda abordagem, aplicou-se a regressão penalizada LASSO, detalhada na Seção 4.2.4. O modelo logístico penalizado foi ajustado em cada reamostragem, e a escolha do parâmetro de regularização λ foi feita por validação cruzada, considerando dois critérios usuais: λ_{min} e λ_{1se}.

Algorithm 2 Reamostragem com seleção via LASSO

Input: Dados D; número de replicações B; procedimento de reamostragem (*Bootstrap(n)*, *Bootstrap(m)* ou *Subsampling(m)*).
for $b = 1$ to B **do**
    Reamostrar D_b conforme o procedimento escolhido;
    Padronizar variáveis preditoras;
    Ajustar modelo LASSO via validação cruzada;
    Selecionar variáveis com $\beta_j \neq 0$ considerando λ_{min} e λ_{1se};
    Registrar variáveis selecionadas e atualizar frequências;
end for
Output: Frequência de inclusão das variáveis ao longo das B reamostragens.

4.4 Resultados e Discussão

A análise da estabilidade da seleção de variáveis pelos métodos clássicos (*backward*, *forward* e *stepwise*) permitiu identificar um conjunto de preditores recorrentes, apesar das diferenças entre os esquemas de reamostragem. De modo geral, o *Subsampling*(m) apresentou comportamento mais conservador, resultando em modelos mais simples e com maior consistência de seleção. Em contraste, o *Bootstrap*(n), especialmente sob níveis de significância mais flexíveis, levou à inclusão de um número maior de variáveis, incluindo preditores com menor regularidade ao longo das reamostragens. Nas Tabelas 40, 41 e 42, apresentadas no Apêndice C, estão descritas as frequências detalhadas de seleção para cada procedimento.

Entre as variáveis avaliadas, **Hemoglobina** e **Bilirrubina Total** apresentaram elevada estabilidade, com frequências de inclusão superiores a 70% na maior parte dos cenários analisados pelos métodos *stepwise*. As exceções ocorreram apenas nas condições mais restritivas de *Bootstrap*(m) e *Subsampling*(m) com $\alpha = 0{,}001$, nas quais as frequências foram menores. Esse padrão indica que esses dois marcadores se mantêm relevantes mesmo sob critérios mais rigorosos de seleção.

Esse comportamento é coerente com a literatura sobre estabilidade na seleção de variáveis, segundo a qual a frequência de inclusão ao longo de reamostragens sucessivas pode ser interpretada como evidência da relevância estrutural do preditor. Variáveis que permanecem selecionadas sob perturbações da amostra tendem a refletir associações mais consistentes, enquanto aquelas com inclusão intermitente sugerem maior dependência de flutuações amostrais (SAUERBREI; BOULESTEIX; BINDER, 2011; ALTMAN; ANDERSEN, 1989).

Na regressão penalizada, os resultados concordam em parte com os métodos clássicos e acrescentam novos elementos à análise. Conforme apresentado na Tabela 1, no cenário mais conservador, que combina *Subsampling*(m) com o critério λ_{1se}, **Creatinina** (84,8%) e **Idade** (73,2%) apresentaram as maiores frequências de seleção. Em seguida, **Lactato** (69,2%) e **Bilirrubina Total** (67,3%) também exibiram frequências elevadas. Esse conjunto configura o núcleo mais estável sob maior penalização, refletindo a tendência do LASSO de favorecer modelos mais simples e menos sensíveis a variações amostrais.

Quando critérios menos restritivos foram adotados, como λ_{min}, ou quando se utilizaram esquemas de reamostragem mais abrangentes, como *Bootstrap*(n) e *Bootstrap*(m), observou-se a inclusão de um número maior de variáveis. Nesse contexto, destacam-se **Hemoglobina** e **Monócitos Relativos**, além da confirmação dos preditores já identificados como mais estáveis. As diferenças entre λ_{min} e λ_{1se} ilustram o equilíbrio entre ajuste e simplicidade: o primeiro tende a priorizar desempenho preditivo, enquanto o segundo favorece modelos mais estáveis e parcimoniosos (FRIEDMAN, 1991; KWON et al., 2023).

Tabela 1 – Frequência de seleção de variáveis via LASSO em combinação com as três estratégias de reamostragem. As colunas indicam a frequência de seleção usando o critério λ_{min} e λ_{1se}.

Variável	B(n)		B(m)		S(m)	
	λ_{min}	λ_{1se}	λ_{min}	λ_{1se}	λ_{min}	λ_{1se}
Creatinina	0.910	0.837	0.848	0.734	0.917	0.848
Idade	0.967	0.849	0.898	0.723	0.927	0.732
Lactato	0.903	0.798	0.830	0.650	0.887	0.692
Bilirrubina Total	0.992	0.910	0.957	0.756	0.959	0.673
Hemoglobina	0.990	0.902	0.943	0.741	0.935	0.636
Monócitos R	0.904	0.745	0.822	0.608	0.859	0.619
Cálcio Ionizado	0.907	0.680	0.820	0.556	0.797	0.444
HCM	0.497	0.496	0.467	0.400	0.479	0.396
Eosinófilos R	0.920	0.671	0.833	0.529	0.778	0.364
Plaquetas	0.878	0.586	0.782	0.471	0.718	0.319
p50	0.823	0.533	0.722	0.420	0.666	0.317
VCM	0.879	0.516	0.752	0.389	0.673	0.199
RDW	0.837	0.490	0.747	0.394	0.599	0.177
Monócitos A	0.677	0.337	0.604	0.285	0.403	0.152
NLCR	0.881	0.450	0.766	0.350	0.603	0.116
PCR	0.891	0.441	0.761	0.330	0.578	0.094
pH	0.726	0.295	0.597	0.227	0.416	0.089
Glicose	0.829	0.365	0.710	0.288	0.495	0.081
Ureia	0.769	0.317	0.648	0.240	0.447	0.077
Metahemoglobina	0.717	0.269	0.650	0.248	0.352	0.067
DAV R	0.622	0.195	0.540	0.182	0.245	0.055
Saturação	0.735	0.275	0.648	0.235	0.363	0.044
Bastonetes A	0.644	0.177	0.570	0.183	0.246	0.038
Linfócitos A	0.670	0.201	0.588	0.194	0.269	0.036
CO_2 Total	0.619	0.164	0.525	0.157	0.199	0.035
PO_2	0.687	0.190	0.598	0.180	0.261	0.026
CHCM	0.579	0.141	0.513	0.149	0.156	0.026
Potássio	0.710	0.201	0.620	0.190	0.279	0.021
Sódio	0.666	0.188	0.584	0.165	0.241	0.020
Carboxihemoglobina	0.703	0.200	0.633	0.194	0.268	0.019
Cloreto	0.644	0.166	0.555	0.147	0.227	0.018
Leucócitos	0.532	0.105	0.419	0.094	0.171	0.013
Segmentados A	0.573	0.129	0.487	0.129	0.176	0.013
pCO_2	0.582	0.129	0.511	0.122	0.180	0.011
Magnésio	0.658	0.152	0.570	0.137	0.227	0.010
Sexo	0.648	0.125	0.535	0.106	0.209	0.010
Segmentados R	0.556	0.113	0.457	0.101	0.158	0.010
Neutrófilos R	0.611	0.127	0.516	0.128	0.184	0.008

Nota: B(n): Bootstrap(n); B(m): Bootstrap(m); S(m): Subsampling(m); R: Valores Relativos; A: Valores Absolutos; HCM: Hemoglobina Corpuscular Média; p50: Pressão de oxigênio a 50% de saturação da hemoglobina; VCM: Volume Corpuscular Médio; RDW: Amplitude de distribuição dos Eritrócitos; NLCR: Razão Neutrófilo-Linfócito; PCR: Proteína C-Reativa; DAV: ; CO_2: Dióxido de Carbono; PO_2: Pressão Parcial de Oxigênio; pCO_2: Pressão Parcial de Dióxido de Carbono.

Também foi observado um grupo de variáveis com frequências intermediárias de seleção, como **Cálcio Ionizado**, **Eosinófilos Relativos** e **Plaquetas**, especialmente nos cenários com λ_{min}, com valores entre aproximadamente 50% e 70%. Esses preditores apresentam relevância dependente do critério de penalização e podem contribuir de forma complementar quando considerados em conjunto com o núcleo mais estável.

Por outro lado, variáveis como **Leucócitos**, **Segmentados Absolutos e Relativos** e **Pressão Parcial de Dióxido de Carbono** apresentaram redução acentuada das frequências sob λ_{1se}, indicando menor estabilidade quando a simplicidade do modelo é priorizada. Em alguns casos, como **Magnésio** e **Hemoglobina Corpuscular Média**, observaram-se frequências moderadas sob λ_{min}, mas queda consistente sob critérios mais conservadores, sugerindo maior sensibilidade ao nível de penalização.

A comparação entre os métodos clássicos e o LASSO mostra ainda que alguns preditores mantêm estabilidade elevada independentemente da abordagem utilizada. **Hemoglobina** e **Bilirrubina Total** destacam-se nesse aspecto, com altas frequências nos procedimentos *stepwise* e valores consistentemente elevados nos diferentes cenários do LASSO. Essa convergência entre métodos distintos reforça a evidência de que esses biomarcadores compõem um núcleo preditivo consistente.

Em contraste, algumas variáveis mostraram forte dependência do método de seleção e do esquema de reamostragem. O **Volume Corpuscular Médio** ilustra esse comportamento, com frequências variando de 87,9% em *Bootstrap*(n) com λ_{min} a 19,9% em *Subsampling*(m) com λ_{1se}. Essa variação sugere que sua relevância pode estar associada a padrões específicos da amostra, tornando sua inclusão menos estável sob critérios mais rigorosos.

De modo geral, os resultados indicam que a estabilidade da seleção de variáveis depende da combinação entre método de seleção e estratégia de reamostragem. O *Subsampling*(m) mostrou-se mais conservador, enquanto o *Bootstrap*(n) refletiu com maior intensidade particularidades da amostra original. A regressão LASSO apresentou comportamento globalmente mais estável e parcimonioso em comparação aos métodos *stepwise*, o que reforça sua adequação em contextos clínicos com múltiplos preditores correlacionados e tamanho amostral moderado.

Esses achados destacam a importância de avaliar empiricamente a estabilidade na seleção de variáveis e de combinar técnicas de reamostragem com métodos de regularização, contribuindo para identificar preditores mais consistentes e reduzir a influência de variações amostrais.

4.5 Conclusão

A seleção de variáveis baseada em estabilidade permitiu identificar um conjunto consistente de biomarcadores com potencial discriminatório para a diferenciação entre bacteremias Gram-positivas e Gram-negativas. De modo geral, a combinação entre reamostragem e regularização favoreceu modelos mais parcimoniosos, com desempenho preditivo comparável entre diferentes configurações analíticas.

Observou-se que a informação diagnóstica relevante tende a concentrar-se em um subconjunto reduzido de preditores, reforçando a utilidade de abordagens multivariáveis na integração de biomarcadores laboratoriais de rotina. Esses achados fundamentam as análises comparativas e a discussão apresentadas no capítulo seguinte.

Capítulo 5

Avaliação de Modelos Preditivos com Conjuntos de Variáveis Estáveis

5.1 Introdução

No Capítulo 4 foi investigada, de forma sistemática, a estabilidade da seleção de variáveis e o impacto de diferentes combinações entre métodos de seleção e esquemas de reamostragem na identificação de preditores robustos para a discriminação entre bacteremias causadas por bactérias Gram-positivas e Gram-negativas. As análises baseadas em reamostragem e regressão penalizada evidenciaram diferenças na consistência dos preditores selecionados e permitiram definir conjuntos de variáveis com distintos níveis de estabilidade e parcimônia.

Neste capítulo, esses conjuntos derivados da análise de estabilidade foram utilizados na etapa de modelagem preditiva. O objetivo é avaliar em que medida o uso de preditores estáveis contribui para a construção de modelos com bom desempenho discriminatório, menor risco de sobreajuste e maior interpretabilidade. Para isso, foram considerados algoritmos de diferentes naturezas, incluindo métodos baseados em árvores, modelos de margem e modelos lineares, permitindo comparar diferentes compromissos entre flexibilidade, simplicidade e capacidade preditiva em dados laboratoriais de rotina.

Especificamente, foram avaliados modelos ajustados com o conjunto completo de variáveis e com três conjuntos derivados da análise de estabilidade: um conjunto estável mais amplo (SMMIN), um conjunto estável mais parcimonioso (SM1SE) e um conjunto mínimo obtido por métodos sequenciais clássicos. O desempenho foi comparado por meio da área sob a curva ROC e métricas complementares, e a relevância relativa dos biomarcadores foi investigada por meio da Medida de Importância por Permutação. Ao final, foram discutidos os modelos que melhor conciliam desempenho discriminatório, parcimônia e interpretabilidade, à luz do objetivo de apoiar a diferenciação entre bacteremias Gram-positivas e Gram-negativas.

5.2 Fundamentação Teórica

5.2.1 Modelos preditivos em contexto clínico: bacteremia

A bacteremia é um evento clínico agudo relevante em serviços de urgência e terapia intensiva, pois sua identificação precoce influencia diretamente decisões como a coleta de hemoculturas e o início da antibioticoterapia empírica. Embora represente apenas uma parte dos quadros infecciosos atendidos nesses cenários, está associada a maior gravidade clínica e a piores desfechos quando não reconhecida a tempo, podendo evoluir rapidamente para sepse (BEN-HAIM et al., 2024; CHIU et al., 2025).

Do ponto de vista metodológico, predizer bacteremia é desafiador. Os sinais clínicos iniciais são frequentemente inespecíficos, os exames microbiológicos demoram para fornecer resultados e os biomarcadores laboratoriais refletem processos inflamatórios complexos. Na prática clínica, a estratificação de risco costuma basear-se em escores como SIRS (Systemic Inflammatory Response Syndrome), SOFA (Sequential Organ Failure Assessment), qSOFA (quick SOFA) e NEWS (National Early Warning Score). Esses instrumentos são simples e amplamente utilizados, mas apresentam capacidade discriminatória limitada, especialmente fora dos contextos em que foram desenvolvidos (FLEUREN et al., 2020; YADGAROV et al., 2024).

Nesse cenário, modelos multivariáveis baseados em dados laboratoriais de rotina têm sido investigados como alternativa. Estudos recentes mostram que algoritmos como *Random Forest*, *Gradient Boosting* e redes neurais podem utilizar informações disponíveis já na triagem para identificar bacteremia, ampliando o potencial preditivo em relação às abordagens tradicionais (MAHMOUD et al., 2021; ZHANG et al., 2023; RATZINGER et al., 2015).

Para a discriminação entre bacteremias Gram-positivas e Gram-negativas, pesquisas indicam que parâmetros hematológicos derivados do hemograma automatizado e dados de população celular podem fornecer informação preditiva relevante (CHIU et al., 2025; DAMBROSO-ALTAFINI et al., 2022). Esses biomarcadores apresentam vantagens práticas importantes: são amplamente disponíveis, de baixo custo e obtidos rapidamente, características desejáveis para ferramentas de apoio à decisão clínica (ZHANG et al., 2025).

Apesar desse potencial, a aplicação clínica desses modelos ainda enfrenta desafios relacionados à interpretabilidade, à estabilidade frente a variações amostrais e ao risco de sobreajuste (*overfitting*), especialmente em contextos com desfechos desbalanceados. Além disso, modelos muito complexos podem dificultar a validação externa e sua incorporação na prática clínica. Por isso, torna-se importante adotar estratégias que busquem equilibrar desempenho preditivo, simplicidade e robustez, motivando o uso de procedimentos de seleção de variáveis baseados em estabilidade, conforme discutido a seguir (MOOR et al., 2021; GARCÍA-LUQUE et al., 2025).

5.2.2 Modelos em Conjunto e Otimização

No aprendizado estatístico, algumas abordagens combinam vários modelos simples para produzir uma predição final mais estável e precisa. Essa estratégia, conhecida como *ensemble learning*, baseia-se na ideia de que modelos diferentes tendem a cometer erros distintos. Ao combiná-los, esses erros podem se compensar, reduzindo a variabilidade e aumentando a robustez das predições. Entre as principais abordagens destacam-se o *Bagging* e o *Boosting*.

O *Bagging* (Bootstrap Aggregating), proposto por Breiman (1996a), foi desenvolvido para reduzir a variância de preditores instáveis. A técnica consiste em ajustar várias versões do mesmo modelo de forma independente, cada uma treinada em uma amostra *bootstrap* do conjunto de dados original. As predições são então combinadas por média (regressão) ou votação (classificação). Esse processo tende a suavizar oscilações e reduzir o risco de sobreajuste. O *Random Forest* é o exemplo mais conhecido desse princípio.

O *Boosting*, introduzido formalmente por Freund e Schapire (1997), segue uma lógica sequencial. Cada novo modelo é ajustado para corrigir os erros cometidos pelos anteriores, concentrando maior atenção nas observações mais difíceis de classificar. Ao final, os modelos são combinados de forma ponderada, formando um preditor mais robusto.

Uma das implementações mais difundidas dessa estratégia é o *Gradient Boosting*, no qual o modelo é construído de forma aditiva, com cada nova árvore ajustada para corrigir os erros acumulados até o momento. Esse ajuste é realizado pela minimização de uma função de perda, utilizando o gradiente como direção de atualização, permitindo ao modelo aprender padrões progressivamente mais complexos (FRIEDMAN, 2001). Implementações modernas baseadas em árvores, como *XGBoost*, *LightGBM* e *HistGradientBoosting*, derivam desse princípio e foram desenvolvidas para melhorar eficiência computacional e desempenho preditivo.

5.2.3 Algoritmos de Classificação

Foram considerados métodos de diferentes paradigmas de aprendizado supervisionado, baseados em árvores de decisão, Máquinas de Vetores de Suporte (SVM) e modelos lineares, de modo a explorar diferentes níveis de flexibilidade, interpretabilidade e desempenho preditivo em dados clínicos.

5.2.3.1 Modelos Baseados em Árvore de Decisão

a) Random Forest

Random Forest (RF), proposto por Breiman (2001), é um método baseado em *Bagging* aplicado a árvores de decisão. O algoritmo constrói múltiplas árvores de forma independente, cada uma treinada a partir de uma amostra *bootstrap* dos dados. Em cada divisão da árvore,

apenas um subconjunto aleatório de variáveis é considerado. Essa aleatoriedade reduz a correlação entre as árvores e resulta em um modelo mais estável e menos suscetível ao sobreajuste. A predição final é obtida por votação majoritária na classificação ou pela média das predições na regressão. O método é capaz de capturar relações não lineares, é robusto a ruído e permite estimar a importância das variáveis, características relevantes em dados clínicos heterogêneos.

b) XGBoost

XGBoost (*Extreme Gradient Boosting*), proposto por Chen e Guestrin (2016), é uma implementação eficiente do *Gradient Boosting* baseada em árvores de decisão. Os modelos são construídos de forma sequencial, com cada nova árvore buscando corrigir os erros das anteriores. O algoritmo incorpora mecanismos de regularização que controlam a complexidade do modelo e reduzem o risco de sobreajuste, contribuindo para boa capacidade de generalização e alto desempenho em dados tabulares.

c) LightGBM

LightGBM (*Light Gradient Boosting Machine*), desenvolvido por Ke et al. (2017), é uma implementação do *Gradient Boosting* otimizada para eficiência computacional e alta dimensionalidade. O algoritmo utiliza crescimento do tipo *leaf-wise*, priorizando divisões que produzem maior redução da função de perda. Essa estratégia tende a melhorar o desempenho preditivo mantendo custo computacional reduzido, o que o torna adequado para conjuntos de dados com muitas variáveis.

d) HistGradientBoosting

Histogram-based Gradient Boosting (*HistGradientBoosting*) é a implementação baseada em histogramas disponível no ecossistema do *scikit-learn* (PEDREGOSA et al., 2011). O algoritmo discretiza variáveis contínuas em intervalos (*bins*) e avalia divisões apenas nesses limites, reduzindo o custo computacional do treinamento. A abordagem é inspirada no *LightGBM*, compartilha o crescimento *leaf-wise* e permite o tratamento nativo de valores ausentes. Sua principal vantagem é a integração direta com as ferramentas de modelagem do *scikit-learn*.

5.2.3.2 Support Vector Machine

Support Vector Machine (SVM), proposto em sua forma moderna por Cortes e Vapnik (1995), é um algoritmo de classificação baseado em margens. Seu objetivo é encontrar o hiperplano que melhor separa as classes no espaço de características, maximizando a distância entre a fronteira de decisão e as observações mais próximas, denominadas *support vectors*.

Quando as classes não são linearmente separáveis, o SVM utiliza o *kernel trick*, que permite representar os dados em um espaço de maior dimensionalidade no qual a separação linear se torna possível. O método é sensível à escala das variáveis, sendo necessária a padronização prévia dos dados. Embora não seja um método de combinação de modelos, o SVM apresenta desempenho competitivo e é frequentemente utilizado como referência em estudos comparativos.

5.2.3.3 Regressão Logística

A regressão logística, formalizada por Cox (1958) é amplamente utilizada para classificação binária em contextos clínicos devido à sua simplicidade, interpretabilidade e base estatística consolidada. Embora seja fundamentalmente um método estatístico clássico, ela é considerada um algoritmo de aprendizado supervisionado linear, servindo frequentemente como modelo de referência em estudos comparativos. O modelo estima a probabilidade de ocorrência do evento de interesse a partir de uma combinação linear dos preditores.

Na prática clínica, a regressão logística é frequentemente adotada como modelo de referência, pois oferece um equilíbrio entre desempenho discriminatório e transparência. Segundo Christodoulou et al. (2019), quando bem conduzida pode apresentar desempenho comparável ao de algoritmos mais complexos de aprendizado estatístico. Seus coeficientes permitem interpretar diretamente a direção e a magnitude da associação entre cada biomarcador e o desfecho, característica importante quando a plausibilidade clínica dos resultados é essencial (HARRELL, 2015; STEYERBERG, 2019). Por ser sensível à escala dos preditores e não admitir valores ausentes, sua aplicação requer procedimentos adequados de pré-processamento, como imputação e padronização das variáveis.

5.2.4 Seleção de variáveis, estabilidade e parcimônia

A seleção de variáveis é uma etapa central no desenvolvimento de modelos preditivos clínicos, pois influencia diretamente o desempenho, a interpretabilidade e a viabilidade de aplicação do modelo. Em estudos baseados em biomarcadores laboratoriais de rotina, o conjunto inicial de preditores costuma ser amplo e frequentemente inclui variáveis correlacionadas ou com contribuição preditiva limitada. Procedimentos adequados de seleção permitem reduzir esse conjunto a um subconjunto mais informativo, diminuindo o risco de sobreajuste e favorecendo modelos mais robustos e aplicáveis (STEYERBERG, 2019; CHOWDHURY; TURIN, 2020).

Diversas estratégias de seleção são descritas na literatura, incluindo métodos baseados em associações univariadas, procedimentos sequenciais (*forward*, *backward* e *stepwise*) e técnicas de regressão penalizada, como LASSO e *elastic net* (TIBSHIRANI, 1996; ZOU; HASTIE, 2005). Em modelos baseados em árvores de decisão, a seleção ocorre de forma implícita, pois

apenas variáveis consideradas informativas são utilizadas nas divisões. No entanto, evidências empíricas indicam que, em cenários com amostras moderadas, desfechos raros e alta correlação entre variáveis, estratégias agressivas de seleção podem gerar modelos instáveis, nos quais pequenas variações nos dados levam à escolha de diferentes conjuntos de preditores (HEINZE; WALLISCH; DUNKLER, 2018; RILEY; COLLINS, 2023).

Por esse motivo, a estabilidade da seleção de variáveis tem sido reconhecida como um critério importante para a confiabilidade de modelos preditivos clínicos. Em vez de depender de um único ajuste, recomenda-se avaliar a consistência da seleção ao longo de múltiplas reamostragens do conjunto de dados, obtidas por meio de técnicas como *bootstrap* ou *subsampling*. Métricas baseadas na frequência de inclusão permitem identificar preditores sistematicamente relevantes e distinguir aqueles mais sensíveis à variação amostral (NOGUEIRA; SECHIDIS; BROWN, 2018; WALLISCH et al., 2021).

Nesse contexto, a *stability selection*, proposta por Meinshausen e Bühlmann (2010), combina regressão penalizada com reamostragem, aumentando a robustez da seleção e reduzindo a inclusão de variáveis espúrias. Extensões desse arcabouço reforçam a utilidade de abordagens baseadas em conjuntos de seletores em cenários com múltiplos preditores (YIN; LI; ZHANG, 2017; MAHMOUDIAN et al., 2021).

A parcimônia do modelo está diretamente relacionada à estabilidade e à interpretabilidade. Modelos muito complexos tendem a apresentar maior variabilidade de desempenho entre amostras e maior risco de sobreajuste, além de dificultarem a interpretação clínica. Por outro lado, modelos excessivamente simples podem comprometer a capacidade discriminatória. Evidências indicam que, em muitos contextos clínicos, é possível alcançar desempenho semelhante com um número reduzido de preditores, desde que selecionados de forma criteriosa e validados adequadamente (RILEY; COLLINS, 2023; DIAZ-RAMIREZ et al., 2021).

Em algoritmos de aprendizado estatístico nos quais a seleção explícita nem sempre é controlada diretamente, medidas de importância global, como a Importância por Permutação, podem ser combinadas com reamostragem para avaliar simultaneamente a relevância e a estabilidade dos preditores (BREIMAN, 2001; FISHER; RUDIN; DOMINICI, 2019). Essa abordagem contribui para identificar variáveis com contribuição consistente para o desempenho do modelo e está alinhada ao objetivo de construir modelos mais simples, interpretáveis e clinicamente aplicáveis.

5.2.5 Otimização de hiperparâmetros

Em algoritmos de aprendizado estatístico, os hiperparâmetros são definidos antes do treinamento e controlam o comportamento global do modelo, diferentemente dos parâmetros, que são estimados diretamente a partir dos dados (AGRAWAL, 2021; HERTEL et al., 2020).

Esses valores regulam aspectos como complexidade do modelo e intensidade de regularização, influenciando o equilíbrio entre viés e variância e, consequentemente, a capacidade de generalização (HASTIE; TIBSHIRANI; FRIEDMAN, 2009; HUTTER; KOTTHOFF; VANSCHOREN, 2019). Por isso, sua definição adequada é uma etapa importante no desenvolvimento de modelos preditivos (AKIBA et al., 2019).

Em modelos baseados em árvores, os hiperparâmetros controlam principalmente a complexidade estrutural e o grau de regularização, influenciando diretamente o risco de sobreajuste (BREIMAN, 2001). Parâmetros como profundidade máxima, número mínimo de observações por nó e número de árvores afetam a estabilidade e a capacidade de generalização do modelo (GEURTS; ERNST; WEHENKEL, 2006). Em métodos de *Boosting*, a taxa de aprendizado regula a contribuição de cada árvore para o modelo final (FRIEDMAN, 2001; HASTIE; TIBSHIRANI; FRIEDMAN, 2009; CHEN; GUESTRIN, 2016). No *Support Vector Machine*, os hiperparâmetros definem a forma da fronteira de decisão, incluindo o parâmetro de regularização C e a escolha do *kernel*. Na regressão logística, hiperparâmetros associados à regularização, como o tipo e a intensidade da penalização, controlam a complexidade do modelo e influenciam sua capacidade de generalização (CORTES; VAPNIK, 1995; HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

A otimização de hiperparâmetros permite explorar o espaço de configurações possíveis de forma estruturada, reduzindo o risco de modelos sobreajustados ou com desempenho limitado. Optuna é uma ferramenta moderna de otimização que utiliza configuração dinâmica do espaço de busca (*define-by-run*) (AKIBA et al., 2019). O método baseia-se no algoritmo TPE (*Tree-structured Parzen Estimator*), que direciona a busca para regiões mais promissoras com base nos resultados obtidos em iterações anteriores (BERGSTRA et al., 2011).

Essa estratégia tende a ser mais eficiente do que métodos exaustivos, como o *grid search*, especialmente em espaços de alta dimensionalidade. Evidências recentes também indicam ganhos de eficiência computacional e desempenho preditivo com o uso do Optuna em comparação com abordagens tradicionais (HUTTER; KOTTHOFF; VANSCHOREN, 2019; SHAO et al., 2024).

5.2.6 Medida de Importância por Permutação

A Medida de Importância por Permutação (*Permutation Feature Importance* — PFI) avalia a relevância de um preditor a partir da variação no desempenho do modelo quando os valores dessa variável são embaralhados aleatoriamente, quebrando sua associação com o desfecho. A importância é quantificada pela redução observada em uma métrica de desempenho, como a ASC, após essa permutação. A ideia foi introduzida por Breiman (2001) no contexto de *Random Forest* e posteriormente formalizada por Fisher, Rudin e Dominici (2019) como uma medida independente do algoritmo.

Uma vantagem importante da PFI em relação às medidas de importância baseadas em impureza é a redução de vieses associados à alta cardinalidade das variáveis. Estudos mostram que métricas baseadas em impureza podem favorecer preditores com muitos valores distintos, mesmo quando sua contribuição preditiva é limitada. Ao avaliar diretamente o impacto da variável no desempenho do modelo, a PFI fornece uma medida mais equilibrada de relevância (ALTMANN et al., 2010; ADLER; PAINSKY, 2022).

Como está diretamente vinculada à métrica de desempenho utilizada, a PFI permite comparar a importância das variáveis entre algoritmos de naturezas distintas, como *Random Forest*, métodos de *Gradient Boosting*, *Support Vector Machine* e regressão logística. Essa característica é particularmente útil em estudos comparativos. Além disso, quando calculada em conjuntos de validação ou teste, a PFI reflete a contribuição das variáveis para a capacidade de generalização do modelo. Estudos indicam que métodos baseados em permutação são eficazes na identificação de variáveis irrelevantes em cenários com ruído ou sobreajuste (DUNN; MINGARDI; ZHUO, 2021; ORLENKO; MOORE, 2021).

Dessa forma, a PFI fornece uma medida global e diretamente interpretável da importância das variáveis, adequada para avaliar a contribuição dos preditores em modelos aplicados a dados clínicos.

5.3 Metodologia

Os dados analisados neste capítulo correspondem ao conjunto previamente descrito e pré-processado no Capítulo 3. Apresentam-se a seguir os procedimentos de organização, particionamento e preparação dos dados para o treinamento, a otimização e a avaliação dos modelos de aprendizado estatístico utilizados na discriminação entre bacteremias Gram-positivas e Gram-negativas. Os experimentos consideram diferentes conjuntos de variáveis definidos a partir dos resultados de estabilidade apresentados no Capítulo 4.

5.3.1 Preparação dos Dados para Modelagem

Os experimentos foram conduzidos na linguagem Python (Python Software Foundation, 2023) utilizando o conjunto de dados descrito no Capítulo 3. A variável resposta foi definida como binária, classificando o agente etiológico como Gram-negativo (1) ou Gram-positivo (0). O conjunto de dados foi dividido aleatoriamente em treinamento (75%) e teste (25%), preservando-se a proporção entre as classes.

A implementação computacional dos modelos de aprendizado de máquina foi realizada na linguagem Python, utilizando a biblioteca *scikit-learn* para a Regressão Logística (função `LogisticRegression`), *Support Vector Machine* (função `SVC`), *Random Forest* (função

`RandomForestClassifier`) e *HistGradientBoosting* (função `HistGradientBoostingClassifier`). Para os modelos baseados em *gradient boosting* otimizados, foram utilizadas as bibliotecas *xgboost* (função `XGBClassifier`) e *lightgbm* (função `LGBMClassifier`). A estruturação e manipulação dos dados ocorreram por meio da biblioteca *pandas* (MCKINNEY et al., 2011).

As etapas de preparação foram definidas de acordo com os requisitos de cada algoritmo, evitando transformações desnecessárias. Nos algoritmos baseados em árvores, não foram realizadas padronização nem transformações adicionais, uma vez que esses métodos são invariantes à escala dos preditores. No caso do *Random Forest*, realizou-se imputação simples pela mediana, pois o algoritmo não admite valores ausentes. Já os métodos *XGBoost*, *LightGBM* e *HistGradientBoosting* tratam valores ausentes nativamente.

Para os modelos *Support Vector Machine* e regressão logística, foi adotado um *pipeline* comum de pré-processamento, uma vez que ambos são sensíveis à escala dos preditores e não admitem valores ausentes. O procedimento incluiu: (i) imputação de valores ausentes pela mediana; (ii) padronização das variáveis numéricas por meio do *StandardScaler*; e (iii) codificação *one-hot* das variáveis categóricas. O ajuste desse *pipeline* foi realizado exclusivamente no conjunto de treinamento, sendo posteriormente aplicado ao conjunto de teste, a fim de evitar vazamento de informação.

5.3.2 Cenários Avaliados

Foram avaliados quatro conjuntos de variáveis para cada algoritmo, totalizando 24 modelos:

- **Modelo completo**: inclui as 38 variáveis disponíveis no conjunto original e serve como referência para avaliar o impacto da redução do número de preditores.
- **SMMIN (11 variáveis)**: conjunto estável obtido por regressão LASSO combinada ao procedimento de *Subsampling*(m), considerando λ_{min}, conforme descrito no Capítulo 4. Inclui: **Creatinina**, **Idade**, **Lactato**, **Bilirrubina Total**, **Hemoglobina**, **Monócitos**, **Cálcio ionizado**, **Eosinófilos**, **Plaquetas**, **P50** e **Volume Corpuscular Médio**.
- **SM1SE (4 variáveis)**: conjunto parcimonioso obtido pelo mesmo procedimento com o critério λ_{1se}, composto por **Creatinina**, **Idade**, **Lactato** e **Bilirrubina Total**.
- **Stepwise (2 variáveis)**: conjunto mínimo formado por **Hemoglobina** e **Bilirrubina Total**, selecionadas de forma consistente pelos métodos clássicos de seleção.

Esses cenários representam diferentes níveis de complexidade e parcimônia, permitindo avaliar o impacto da redução do número de preditores no desempenho discriminatório e no equilíbrio entre simplicidade, estabilidade e capacidade preditiva.

5.3.3 Treinamento e Otimização dos Modelos

Para cada combinação de algoritmo e conjunto de variáveis, realizou-se otimização automática de hiperparâmetros com Optuna, utilizando o algoritmo TPE (*Tree-structured Parzen Estimator*). A função objetivo foi a ASC-ROC, estimada por validação cruzada estratificada com 10 partições no conjunto de treinamento.

Os espaços de busca foram definidos com base na literatura metodológica e em recomendações dos trabalhos seminais dos algoritmos, bem como em práticas consolidadas das bibliotecas *scikit-learn*, *XGBoost* e *LightGBM* (HASTIE; TIBSHIRANI; FRIEDMAN, 2009; HARRELL, 2015; STEYERBERG, 2019; BREIMAN, 2001; CHEN; GUESTRIN, 2016; KE et al., 2017; AKIBA et al., 2019). Foram realizadas 1000 iterações para cada modelo.

Após a identificação da melhor configuração média na validação cruzada, o modelo final foi ajustado utilizando todo o conjunto de treinamento com os hiperparâmetros selecionados. A avaliação final foi realizada exclusivamente no conjunto de teste, garantindo estimativa independente do desempenho de generalização.

5.3.4 Métricas de Avaliação

O desempenho foi avaliado por métricas complementares de discriminação:

- **ASC-ROC** (Área sob a Curva ROC), utilizada como métrica principal, por ser independente do limiar de classificação;
- **Acurácia**, definida como a proporção de predições corretas;
- **Sensibilidade**, que mede a capacidade de identificar corretamente os casos positivos;
- **Especificidade**, que avalia a identificação correta dos casos negativos.

Os intervalos de confiança de 95% da ASC-ROC foram estimados por *bootstrap* no conjunto de teste (2000 reamostragens). A avaliação foi complementada por matriz de confusão e curva ROC.

A interpretabilidade foi investigada por meio da Importância por Permutação, que quantifica a contribuição de cada variável para o desempenho do modelo e permite comparação consistente entre algoritmos distintos.

5.4 Resultados e Discussão

A Tabela 2 resume o desempenho no conjunto de teste em termos de ASC-ROC, acurácia, sensibilidade e especificidade. Observa-se variação da ASC-ROC conforme a combinação

entre algoritmo e conjunto de variáveis, sem evidência de superioridade consistente de um algoritmo em todos os cenários. De modo geral, os resultados indicam que a estratégia de seleção de variáveis exerce influência comparável, ou superior, à escolha do algoritmo.

Tabela 2 – Desempenho dos modelos de classificação no conjunto de teste, considerando diferentes algoritmos e métodos de seleção de variáveis.

Algoritmo	**Metodo de Seleção**	**ASC**	**Acurácia**	**Sensibilidade**	**Especificidade**
RL	Completo	0,6418	0,552	0,458	0,636
	SMMIN	**0,7527**	0,656	0,559	0,742
	SM1SE	0,7218	0,672	0,712	0,636
	Stepwise	0,5986	0,544	0,339	0,727
RF	Completo	0,6998	0,664	0,508	0,803
	SMMIN	**0,7406**	0,656	0,491	0,803
	SM1SE	0,6795	0,600	0,390	0,788
	Stepwise	0,5643	0,528	0,356	0,682
XGB	Completo	0,6184	0,576	0,491	0,651
	SMMIN	**0,6965**	0,616	0,491	0,727
	SM1SE	0,6177	0,600	0,441	0,742
	Stepwise	0,5517	0,512	0,339	0,667
LGBM	Completo	0,6592	0,632	0,559	0,697
	SMMIN	**0,6613**	0,648	0,542	0,742
	SM1SE	0,6608	0,648	0,508	0,773
	Stepwise	0,5116	0,456	0,288	0,606
HGB	Completo	0,6531	0,648	0,559	0,727
	SMMIN	**0,6718**	0,616	0,475	0,742
	SM1SE	0,6319	0,616	0,491	0,727
	Stepwise	0,5235	0,464	0,288	0,621
SVM	Completo	0,6538	0,576	0,458	0,682
	SMMIN	0,6492	0,600	0,525	0,667
	SM1SE	**0,6879**	0,656	0,797	0,530
	Stepwise	0,5833	0,544	0,305	0,758

Nota: RL: Regressão Logística; RF: *Random Forest*; XGB: XGBoost; LGBM: LightGBM; HGB: HistGradientBoosting; SVM: *Support Vector Machine*. Os métodos de seleção e o número de variáveis: *Stepwise* (2), SM1SE (4), SMMIN (11) e Completo (38). ASC: área sob a curva ROC. Valores em negrito indicam o maior valor de ASC-ROC para cada algoritmo.

O desempenho competitivo da regressão logística, particularmente quando combinada com conjuntos de variáveis estáveis, é consistente com evidências recentes da literatura. Revisão sistemática conduzida por Christodoulou et al. (2019) não identificou superioridade consistente de algoritmos de aprendizado de máquina complexos sobre a regressão logística em dados clínicos estruturados. Resultados semelhantes foram relatados por Song et al. (2021), que observaram desempenho comparável entre regressão logística e métodos de aprendizado estatístico em dados tabulares de dimensão moderada.

Entre os cenários avaliados, o conjunto SMMIN (11 variáveis) apresentou, em geral, os maiores valores de ASC-ROC. Destacam-se a regressão logística (ASC-ROC = 0,7527) e o *Random Forest* (ASC-ROC = 0,7406), ambos com desempenho superior ao obtido com o conjunto completo de 38 variáveis. Esse resultado é compatível com a redução de dimensionalidade associada à seleção baseada em estabilidade, que preserva preditores informativos e reduz a influência de variáveis redundantes ou que pouco contribuem (CUETO-LÓPEZ et al., 2019; YU et al., 2023).

A Figura 2 apresenta os valores de ASC-ROC para cada combinação entre algoritmo e método de seleção. Observa-se concentração de maiores valores nos cenários associados ao SMMIN, enquanto os cenários baseados em *stepwise* tendem a apresentar desempenho inferior. Esse padrão é consistente com limitações descritas para procedimentos sequenciais, particularmente quanto à instabilidade da seleção de preditores (BREIMAN, 1996b; STEYERBERG, 2019).

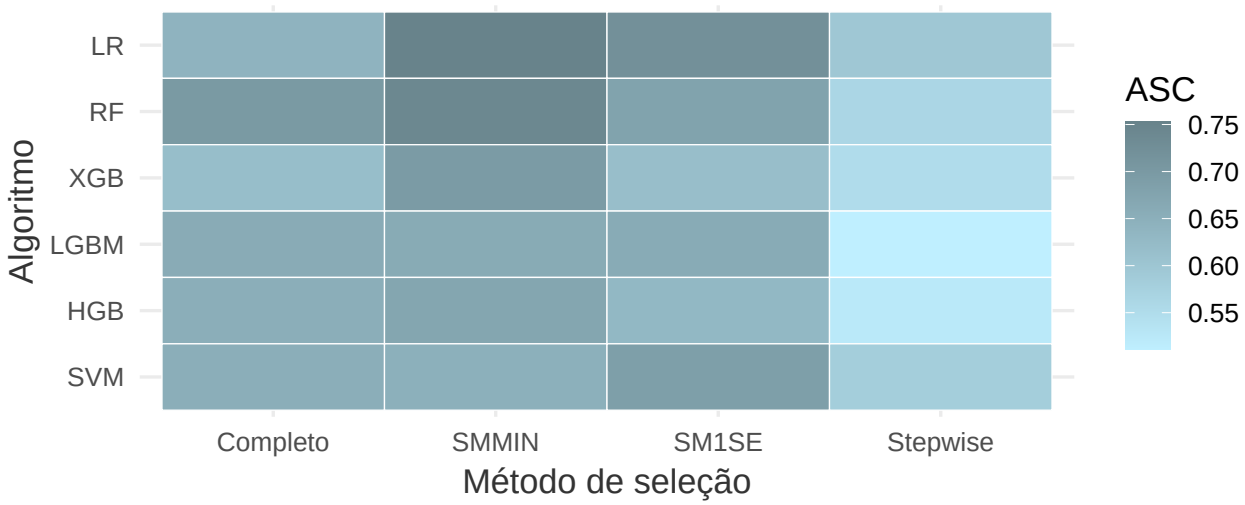

Figura 2 – *Heatmap* da área sob a curva ROC (ASC-ROC) para cada combinação entre algoritmo de classificação e método de seleção de variáveis. Cores mais intensas indicam melhor desempenho discriminatório.

A utilização do conjunto completo de variáveis não se associou a ganhos sistemáticos de desempenho. Em diversos cenários, os valores de ASC-ROC foram semelhantes ou inferiores aos observados após a seleção, sugerindo que a inclusão de grande número de preditores não necessariamente melhora a discriminação e pode aumentar o risco de sobreajuste em amostras de tamanho moderado (HASTIE; TIBSHIRANI; FRIEDMAN, 2009; GARCÍA-LUQUE et al., 2025). Esse resultado reforça a utilidade de modelos mais parcimoniosos em aplicações clínicas.

A Figura 3 mostra a ASC-ROC dos algoritmos em função do método de seleção. A regressão logística apresenta ganhos mais evidentes quando combinada com conjuntos estáveis

(SMMIN e SM1SE), enquanto os algoritmos de conjunto exibem desempenho relativamente estável entre os cenários, possivelmente em razão de sua capacidade de lidar com colinearidade e realizar seleção interna de atributos (BREIMAN, 2001; CHEN; GUESTRIN, 2016). Esse padrão sugere que o desempenho da regressão logística depende mais diretamente da qualidade e da dimensão do conjunto de preditores.

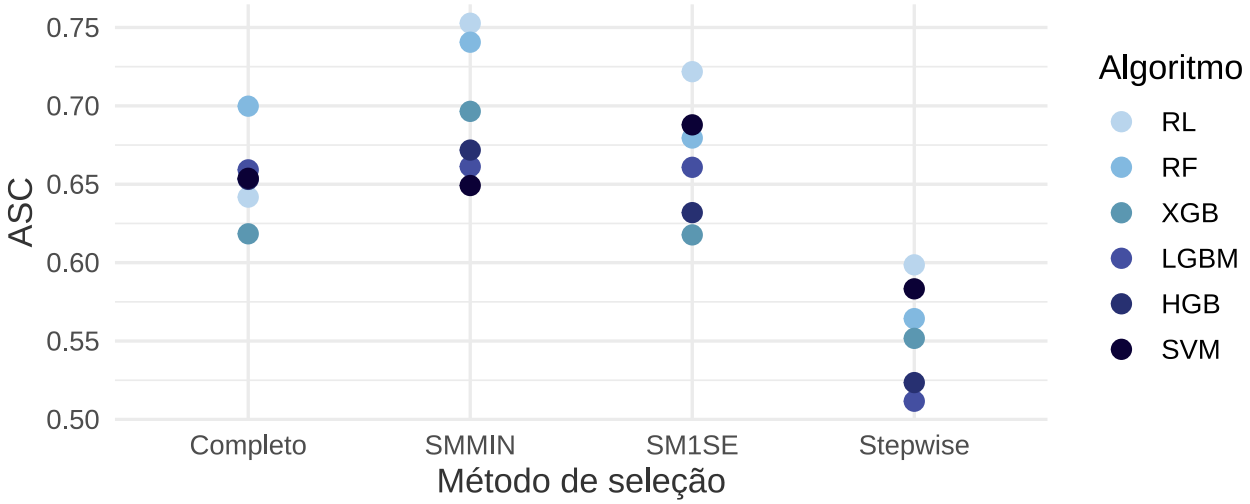

Figura 3 – Área sob a curva ROC (ASC) obtida pelos diferentes algoritmos de classificação em função do método de seleção de variáveis. Cada ponto representa o desempenho de um algoritmo para um determinado método de seleção.

A relação entre desempenho e número de variáveis é apresentada na Figura 4. Conjuntos intermediários apresentaram ASC-ROC média superior à obtida com conjuntos muito reduzidos ou com o modelo completo, indicando um equilíbrio entre parcimônia e desempenho discriminatório.

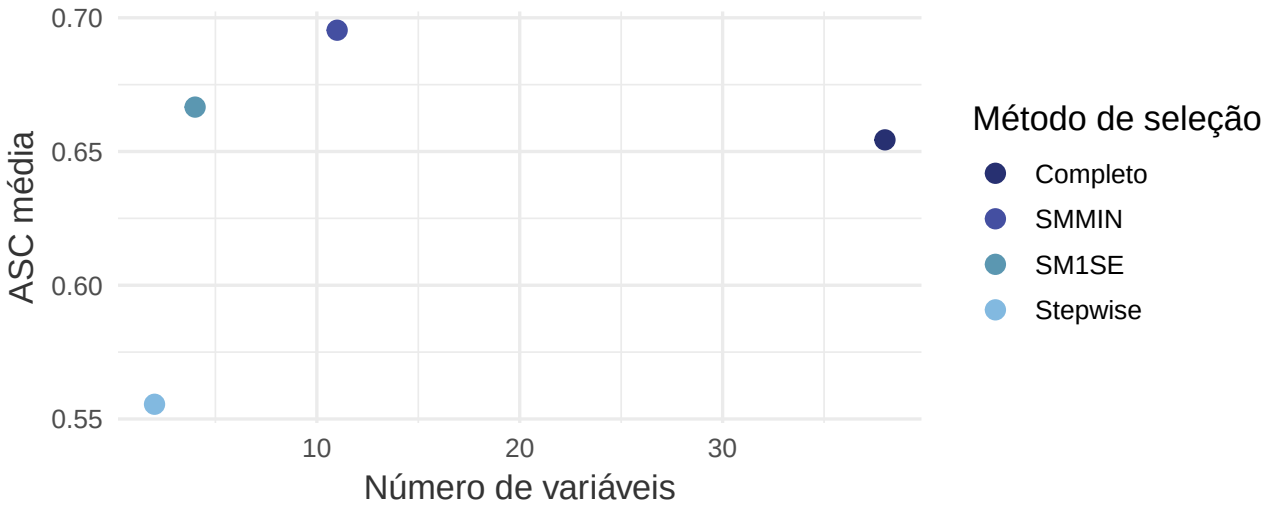

Figura 4 – Relação entre desempenho discriminatório (ASC-ROC média entre algoritmos) e número de variáveis selecionadas para cada método de seleção.

Esse resultado está de acordo com evidências de que, em muitos contextos clínicos, desempenho semelhante pode ser alcançado com número reduzido de preditores adequadamente selecionados (HARRELL, 2015; DIAZ-RAMIREZ et al., 2021; MATHARAARACHCHI; DOMARATZKI; MUTHUKUMARANA, 2022).

De forma geral, os resultados indicam maior impacto da estratégia de seleção de variáveis em comparação com a escolha do algoritmo. Métodos baseados em estabilidade apresentaram desempenho mais consistente, enquanto abordagens sequenciais clássicas mostraram desempenho inferior na maioria dos cenários avaliados.

Em conjunto, os resultados indicam que biomarcadores laboratoriais de rotina apresentam capacidade discriminatória moderada, porém consistente, para apoio à decisão clínica inicial. Observa-se ainda que a qualidade e a estabilidade do conjunto de preditores exercem influência mais direta sobre o desempenho do modelo do que a complexidade do algoritmo, em concordância com evidências comparativas em dados clínicos estruturados (CHRISTODOULOU et al., 2019; WANG et al., 2025).

5.4.1 Modelos Selecionados para Análise Detalhada

Com base nos resultados gerais, foi selecionado um subconjunto de modelos para análise detalhada, priorizando cenários com maior relevância metodológica e potencial aplicabilidade clínica. A seleção considerou três critérios: desempenho discriminatório (ASC-ROC), parcimônia e interpretabilidade.

Embora o modelo de *Random Forest* (RF–SMMIN) tenha apresentado elevada ASC-

ROC, ele não mostrou superioridade consistente nas demais métricas de desempenho, como sensibilidade e especificidade, além de apresentar menor interpretabilidade e maior complexidade estrutural. Dessa forma, optou-se por priorizar modelos que combinassem bom desempenho discriminatório, simplicidade e potencial de aplicação clínica.

Foram selecionados dois modelos de regressão logística baseados em seleção por estabilidade, representando diferentes níveis de complexidade. Os modelos finais foram ajustados com os hiperparâmetros definidos pelo procedimento descrito na Seção 5.3.3.

O primeiro modelo corresponde à regressão logística com seleção SMMIN (RL–SMMIN), composta por 11 variáveis, que apresentou o maior valor de ASC-ROC entre os cenários avaliados (ASC-ROC = 0,753; IC 95%: 0,667–0,833). Esse modelo foi adotado como referência metodológica por representar o melhor desempenho discriminatório observado (MEINSHAUSEN; BÜHLMANN, 2010; NOGUEIRA; SECHIDIS; BROWN, 2018).

O segundo modelo corresponde à regressão logística com seleção SM1SE (RL–SM1SE), composta por quatro variáveis. Embora tenha apresentado ASC-ROC ligeiramente inferior (ASC-ROC = 0,722; IC 95%: 0,634–0,811), exibiu maior sensibilidade e maior simplicidade estrutural, caracterizando um cenário de maior parcimônia com desempenho competitivo.

5.4.2 Desempenho dos Modelos Selecionados

A Figura 5 apresenta as curvas ROC dos modelos RL–SMMIN e RL–SM1SE no conjunto de teste. O modelo RL–SMMIN apresentou maior ASC-ROC, indicando melhor discriminação global na amostra analisada. A sobreposição parcial dos intervalos de confiança sugere, contudo, interpretação cautelosa da diferença entre os modelos. O desempenho do RL–SM1SE permanece competitivo, apesar do menor número de preditores.

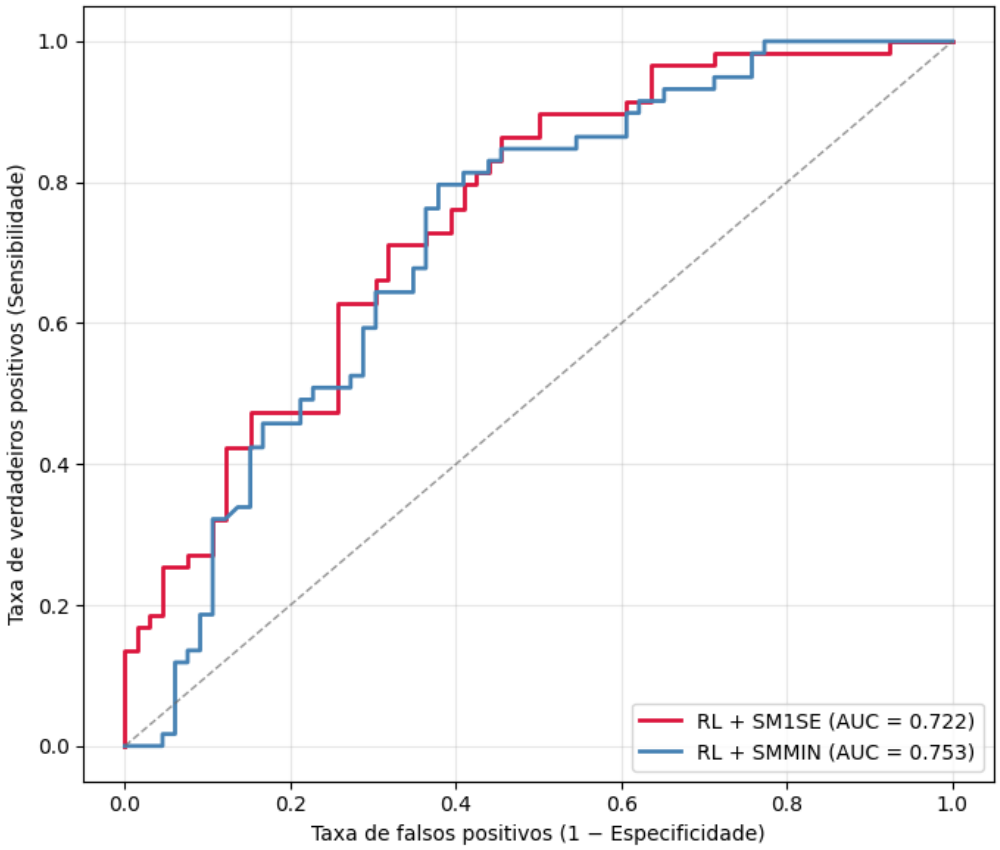

Figura 5 – Curvas ROC dos modelos de regressão logística (RL) com seleção SMMIN e SM1SE no conjunto de teste.

As matrizes de confusão (Figura 6) evidenciam perfis distintos de classificação. O modelo RL–SM1SE apresenta maior sensibilidade, com aumento de verdadeiros positivos ao custo de redução da especificidade. O RL–SMMIN apresenta equilíbrio mais uniforme entre sensibilidade e especificidade, coerente com seu maior valor global de ASC-ROC.

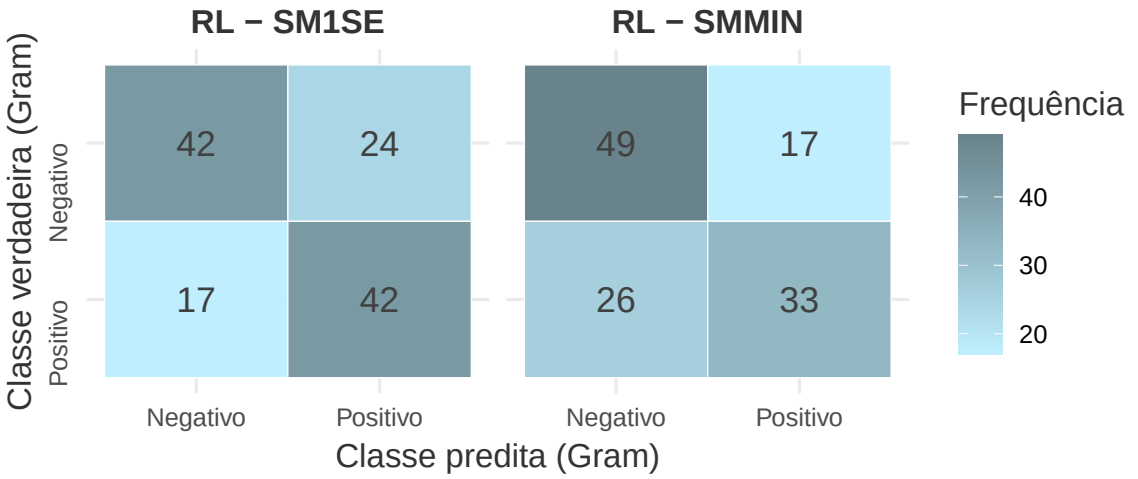

Figura 6 – Matrizes de confusão dos modelos de regressão logística (RL) com seleção SMMIN e SM1SE no conjunto de teste.

A importância das variáveis foi avaliada por meio da Importância por Permutação. A Figura 7 apresenta a redução média da ASC após a permutação de cada preditor. No modelo RL–SM1SE, **Creatinina**, **Idade** e **Lactato** apresentaram maior contribuição para o desempenho, enquanto **Bilirrubina Total** apresentou contribuição menor. No modelo RL–SMMIN, **Hemoglobina**, **Idade** e **Lactato** apresentaram maior importância relativa, enquanto as demais variáveis exibiram contribuições mais discretas, possivelmente complementares no contexto multivariável.

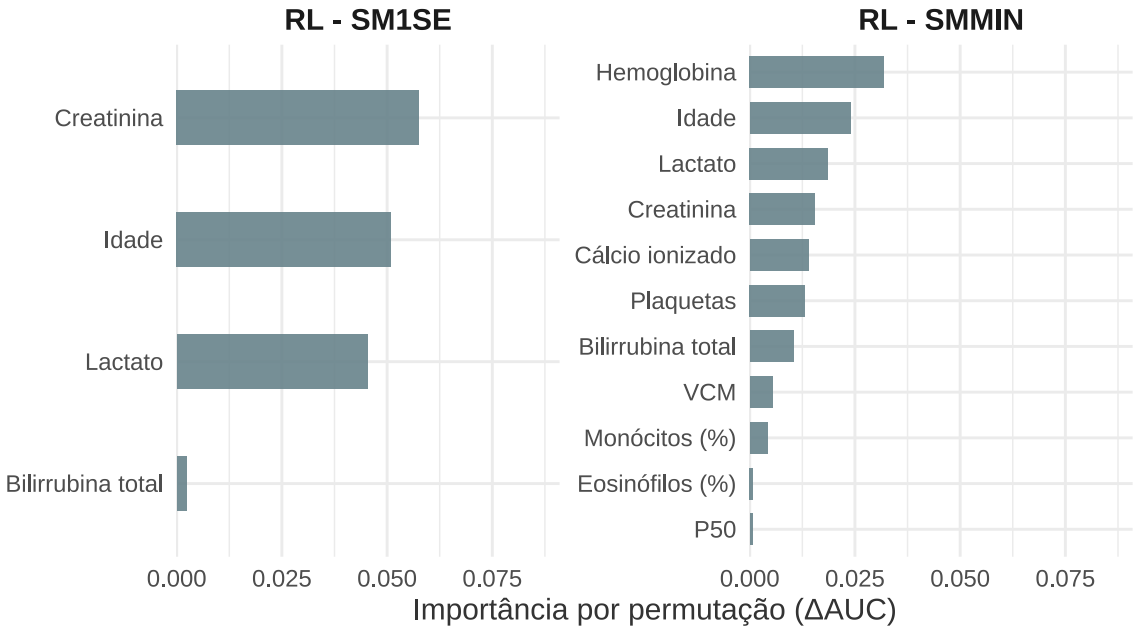

Figura 7 – Importância das variáveis estimada pelo método de permutação (PFI) para os modelos RL–SM1SE e RL–SMMIN.

De forma geral, variáveis associadas à disfunção metabólica e inflamatória apresentaram maior importância em ambos os modelos, em concordância com achados descritos na literatura sobre biomarcadores laboratoriais e discriminação etiológica em bacteremia (DAMBROSO-ALTAFINI et al., 2022; MURRI et al., 2024; ZHANG et al., 2023).

5.4.3 Comparação com estudos prévios

A contribuição dos modelos propostos torna-se mais clara quando seus resultados são comparados com estudos anteriores. Utilizando a mesma base de dados analisada neste trabalho, Dambroso-Altafini et al. (2022) desenvolveram um modelo de regressão logística com seleção sequencial clássica, resultando em um modelo reduzido com quatro variáveis (Plaquetas, Creatinina, HCM e Hemácias). Ao compará-lo com o modelo parcimonioso deste estudo (RL–SM1SE), também composto por quatro preditores (Creatinina, Idade, Lactato e Bilirrubina Total), observa-se melhora tanto na sensibilidade (0,610 para 0,712) quanto na área sob a curva ROC (0,690 para 0,722). Esse resultado sugere que a utilização de procedimentos baseados em regularização e estabilidade pode contribuir para a construção de modelos mais robustos, mesmo mantendo um número reduzido de variáveis.

Essas diferenças podem ser visualizadas de forma resumida na Tabela 3, que apresenta uma comparação entre resultados da literatura e os obtidos nesta dissertação.

Tabela 3 – Comparação do desempenho preditivo dos modelos propostos com estudos da literatura.

Estudo	**Modelo**	**Nº var**	**Sens.**	**Esp.**	**ASC-ROC**
Ratzinger et al. (2015)	K-Star	7	0,446	N/D	0,675
Dambroso-Altafini et al. (2022)	RL	4	0,610	0,670	0,690
Modelos Propostos					
RL-SM1SE	RL	4	0,712	0,636	0,722
RL-SMMIN	RL	11	0,559	0,742	0,753

Nota: RL = Regressão Logística; Nº var = número de variáveis do modelo; Sens. = Sensibilidade; Esp. = Especificidade; N/D = Não disponível ou não reportado nos resultados do modelo principal.

Alguns estudos anteriores demonstraram cautela quanto à capacidade de biomarcadores laboratoriais de rotina em discriminar a etiologia bacteriana. Ratzinger et al. (2015), por exemplo, aplicaram algoritmos de aprendizado de máquina para diferenciar bacteremias Gram-positivas e Gram-negativas a partir de exames laboratoriais e observaram desempenho limitado para essa tarefa, com área sob a curva ROC de aproximadamente 0,675 e sensibilidade inferior a 50%.

Os resultados obtidos nesta dissertação indicam que esses biomarcadores podem apresentar melhor desempenho quando analisados com estratégias estatísticas voltadas à estabilidade da seleção de variáveis. Nesse contexto, o modelo RL–SMMIN alcançou ASC-ROC de 0,753, enquanto o modelo RL–SM1SE apresentou sensibilidade de 0,712 no conjunto de teste. Além do desempenho preditivo, outro aspecto relevante é a facilidade de obtenção das variáveis utilizadas. Na prática hospitalar, esses biomarcadores podem ser obtidos a partir de exames amplamente solicitados em pacientes com suspeita de infecção grave, como o hemograma completo e a gasometria arterial com parâmetros metabólicos (TANG et al., 2020; LIEN et al., 2022).

Além do desempenho preditivo, outro aspecto relevante é a facilidade de obtenção das variáveis utilizadas. Na prática hospitalar, esses biomarcadores podem ser obtidos a partir de exames amplamente solicitados em pacientes com suspeita de infecção grave, como o hemograma completo e a gasometria arterial com parâmetros metabólicos. No modelo RL–SM1SE, todos os biomarcadores podem ser obtidos a partir de um único exame laboratorial de rotina. Já o modelo RL–SMMIN, embora inclua um número maior de preditores, requer apenas dois exames laboratoriais para a obtenção de todos os biomarcadores necessários. No conjunto total de variáveis utilizadas, a maior parte corresponde a parâmetros obtidos por gasometria arterial, enquanto outra parcela é derivada do hemograma completo (TANG et al., 2020; LIEN et al., 2022; ZHANG et al., 2023).

Esses exames são amplamente disponíveis e apresentam tempo de liberação relativamente rápido (COLAK et al., 2020). A gasometria arterial, em particular, costuma ser processada com prioridade nos laboratórios hospitalares, permitindo a obtenção de resultados em poucos minutos. Dessa forma, os modelos propostos podem contribuir para estimar precocemente a provável etiologia da bacteremia ainda nas primeiras horas de atendimento, antes da confirmação microbiológica por hemocultura, que frequentemente demanda dias para fornecer o patógeno definitivo (PETERS et al., 2004; LOONEN et al., 2014). Além disso, a utilização de procedimentos baseados em estabilidade favorece a seleção de preditores que não são apenas estatisticamente relevantes, mas também clinicamente plausíveis e operacionalmente acessíveis, o que reforça o potencial de aplicação desses modelos em contextos assistenciais reais (ZHANG et al., 2023).

5.5 Conclusão

Neste capítulo, avaliou-se o desempenho de diferentes algoritmos de classificação a partir de conjuntos de variáveis definidos com base na estabilidade da seleção. Os resultados mostraram que a qualidade e a robustez do conjunto de preditores exercem influência mais marcante sobre o desempenho discriminatório do que a complexidade do algoritmo empregado, com des-

taque para a regressão logística ajustada aos conjuntos estáveis.

O conjunto estável ampliado apresentou a melhor capacidade de discriminação entre bacteremias Gram-positivas e Gram-negativas, enquanto o modelo parcimonioso, baseado em um número reduzido de preditores, obteve desempenho competitivo aliado a elevada sensibilidade. Essa combinação de boa acurácia, parcimônia e estabilidade sugere que é possível construir modelos úteis para triagem inicial a partir de poucos biomarcadores.

Outro aspecto relevante é que os preditores utilizados pelos modelos propostos são derivados de exames laboratoriais de rotina, como hemograma e gasometria, que são obtidos rapidamente na prática clínica. Isso reforça o potencial de aplicação prática dos modelos desenvolvidos, tanto em serviços de maior complexidade quanto em contextos com recursos mais limitados.

De forma geral, os resultados deste capítulo indicam que estratégias de seleção de variáveis baseadas em estabilidade, combinadas a modelos estatísticos interpretáveis, podem contribuir para a diferenciação precoce entre bacteremias Gram-positivas e Gram-negativas, apoiando decisões terapêuticas mais rápidas e fundamentadas.

Capítulo 6

Conclusão

Biomarcadores laboratoriais de rotina apresentam informação preditiva relevante para diferenciar bacteremias causadas por bactérias Gram-positivas e Gram-negativas. A análise exploratória inicial indicou diferenças consistentes entre esses grupos, sugerindo que tais marcadores podem ser explorados em modelos multivariáveis de predição.

A análise da estabilidade da seleção de variáveis, baseada na regressão penalizada LASSO e em esquemas de reamostragem como o *Subsampling*, mostrou-se útil para identificar preditores robustos e reduzir a influência de flutuações amostrais. Os modelos de aprendizado estatístico desenvolvidos apresentaram desempenho discriminatório moderado e consistente, com resultados semelhantes entre diferentes algoritmos quando alimentados por conjuntos de variáveis estáveis.

Observou-se que a qualidade e a estabilidade dos preditores tiveram maior impacto no desempenho dos modelos do que a complexidade do algoritmo utilizado. Em particular, o núcleo reduzido de biomarcadores identificado pela análise de estabilidade manteve desempenho competitivo mesmo em modelos mais simples, sendo composto por Creatinina, Idade, Lactato e Bilirrubina Total.

Apesar dos resultados obtidos e do rigor metodológico aplicado na seleção de variáveis, este estudo apresenta algumas limitações que devem ser consideradas, muitas delas relacionadas à própria base de dados utilizada. Trata-se de um estudo retrospectivo e unicêntrico, baseado exclusivamente em dados do Hospital Universitário de Maringá (HUM). Essa característica pode limitar a generalização dos resultados. O desempenho dos modelos pode variar entre instituições com diferentes perfis epidemiológicos, rotinas laboratoriais e prevalência de patógenos.

Além disso, por se tratar de uma base retrospectiva, a coleta de dados está sujeita à ausência de algumas informações clínicas e laboratoriais. Embora a base consolidada tenha incluído 523 pacientes, o desenvolvimento de modelos de aprendizado de máquina frequentemente se beneficia de amostras ainda maiores. Dessa forma, a validação externa dos modelos em co-

ortes prospectivas e multicêntricas representa um passo importante para avaliar sua robustez e aplicabilidade em diferentes contextos clínicos.

Como perspectivas futuras, recomenda-se realizar a validação externa dos modelos em dados provenientes de outros centros hospitalares e atualizar os modelos com novos dados ao longo do tempo. Além disso, futuras investigações podem explorar técnicas para identificar padrões de coocorrência entre biomarcadores, avaliando quais variáveis tendem a ser selecionadas em conjunto em diferentes cenários amostrais.

Os resultados desta dissertação indicam que considerar explicitamente a estabilidade na seleção de variáveis é uma estratégia adequada para o desenvolvimento de modelos preditivos com dados laboratoriais clínicos. Os modelos propostos oferecem uma base quantitativa para apoiar a diferenciação precoce entre bacteremias Gram-positivas e Gram-negativas. Além disso, os resultados mostram que abordagens mais simples, baseadas em preditores estáveis e modelos interpretáveis, podem alcançar bom desempenho preditivo mantendo transparência e potencial de aplicação clínica.

Referências

ABU-EL-RUZ, R. et al. Artificial Intelligence in Bacterial Infections Control: A Scoping Review. *Antibiotics*, v. 14, n. 3, p. 256, 2025. Disponível em: <https://doi.org/10.3390/antibiotics14030256>. Acesso em: 5 mar. 2026.

ADLER, A. I.; PAINSKY, A. Feature Importance in Gradient Boosting Trees with Cross-Validation Feature Selection. *Entropy*, v. 24, n. 5, p. 687, 2022. Disponível em: <https://doi.org/10.3390/e24050687>. Acesso em: 5 mar. 2026.

AGRAWAL, T. *Hyperparameter Optimization in Machine Learning*: Make Your Machine Learning and Deep Learning Models More Efficient. Berkeley: Apress, 2021.

AKAIKE, H. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, v. 19, n. 6, p. 716–723, 1974. Disponível em: <https://doi.org/10.1109/TAC.1974.1100705>. Acesso em: 5 mar. 2026.

AKIBA, T. et al. Optuna: A next-generation hyperparameter optimization framework. In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. New York, NY, USA: ACM, 2019. p. 2623–2631. Disponível em: <https://doi.org/10.1145/3292500.3330701>. Acesso em: 5 mar. 2026.

ALTERTHUM, F. *Microbiologia*. 6. ed. São Paulo: Editora Atheneu, 2015.

ALTMAN, D. G.; ANDERSEN, P. K. Bootstrap investigation of the stability of a Cox regression model. *Statistics in Medicine*, v. 8, n. 7, p. 771–783, 1989. Disponível em: <https://doi.org/10.1002/sim.4780080702>. Acesso em: 5 mar. 2026.

ALTMANN, A. et al. Permutation importance: A corrected feature importance measure. *Bioinformatics*, v. 26, n. 10, p. 1340–1347, 2010. Disponível em: <https://doi.org/10.1093/bioinformatics/btq134>. Acesso em: 5 mar. 2026.

AUSTIN, P. C.; TU, J. V. Bootstrap methods for developing predictive models. *The American Statistician*, v. 58, n. 2, p. 131–137, 2004. Disponível em: <https://doi.org/10.1198/0003130043277>. Acesso em: 5 mar. 2026.

BASSETTI, M. et al. Role of procalcitonin in predicting etiology in bacteremic patients: Report from a large single-center experience. *Journal of Infection and Public Health*, v. 13, n. 1, p. 40–45, 2020. Disponível em: <https://doi.org/10.1016/j.jiph.2019.06.003>. Acesso em: 5 mar. 2026.

BEALE, E. M. L. Note on procedures for variable selection in multiple regression. *Technometrics*, v. 12, n. 4, p. 909–914, 1970. Disponível em: <https://doi.org/10.2307/1267335>. Acesso em: 5 mar. 2026.

BEN-HAIM, G. et al. Combination of machine learning algorithms with natural language processing may increase the probability of bacteremia detection in the emergency department: A retrospective, big-data analysis of 94,482 patients. *Digital Health*, v. 10, p. 1–11, 2024. Disponível em: <https://doi.org/10.1177/20552076241277673>. Acesso em: 5 mar. 2026.

BERGSTRA, J. et al. Algorithms for hyper-parameter optimization. *Advances in Neural Information Processing Systems*, v. 24, p. 2546–2554, 2011. Disponível em: <https://dl.acm.org/doi/abs/10.5555/2986459.2986743>. Acesso em: 5 mar. 2026.

BEYENE, J. et al. Determining relative importance of variables in developing and validating predictive models. *BMC Medical Research Methodology*, v. 9, n. 1, p. 64, 2009. Disponível em: <https://doi.org/10.1186/1471-2288-9-64>. Acesso em: 5 mar. 2026.

BIN, R. D. et al. Subsampling versus bootstrapping in resampling-based model selection for multivariable regression. *Biometrics*, v. 72, n. 1, p. 272–280, 2016. Disponível em: <https://doi.org/10.1111/biom.12381>. Acesso em: 5 mar. 2026.

BOX, G. E. P. Robustness in the strategy of scientific model building. In: *Robustness in Statistics*. New York: Academic Press, 1979. p. 201–236. Disponível em: <https://doi.org/10.1016/B978-0-12-438150-6.50018-2>. Acesso em: 5 mar. 2026.

BREIMAN, L. Bagging predictors. *Machine Learning*, v. 24, n. 2, p. 123–140, 1996. Disponível em: <https://doi.org/10.1007/bf00058655>. Acesso em: 5 mar. 2026.

BREIMAN, L. Heuristics of instability and stabilization in model selection. *The Annals of Statistics*, v. 24, n. 6, p. 2350–2383, 1996. Disponível em: <https://doi.org/10.1214/aos/1032181158>. Acesso em: 5 mar. 2026.

BREIMAN, L. Random Forests. *Machine Learning*, v. 45, n. 1, p. 5–32, 2001. Disponível em: <https://doi.org/10.1023/A:1010933404324>. Acesso em: 5 mar. 2026.

BROUWER, L.; CUNNEY, R.; DREW, R. J. Predicting community acquired bloodstream infection in infants using full blood count parameters and C-reactive protein: A machine learning study. *European Journal of Pediatrics*, v. 183, n. 7, p. 2983–2993, 2024. Disponível em: <https://doi.org/10.1007/s00431-024-05441-6>. Acesso em: 5 mar. 2026.

BURNHAM, K. P.; ANDERSON, D. R. *Model Selection and Multimodel Inference*: A Practical Information-Theoretic Approach. 2. ed. New York: Springer, 2002.

CHEN, C.-H.; GEORGE, S. L. The bootstrap and identification of prognostic factors via Cox's proportional hazards regression model. *Statistics in Medicine*, v. 4, n. 1, p. 39–46, 1985. Disponível em: <https://doi.org/10.1002/sim.4780040107>. Acesso em: 5 mar. 2026.

CHEN, T.; GUESTRIN, C. XGBoost: A scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY, USA: ACM, 2016. p. 785–794. Disponível em: <https://doi.org/10.1145/2939672.2939785>. Acesso em: 5 mar. 2026.

CHIU, Y.-W. et al. Monitoring for early prediction of Gram-negative bacteremia using machine learning and hematological data in the emergency department. *Communications Medicine*, v. 5, n. 1, p. 483, 2025. Disponível em: <https://doi.org/10.1038/s43856-025-01200-2>. Acesso em: 5 mar. 2026.

CHOWDHURY, M. Z. I.; TURIN, T. C. Variable selection strategies and its importance in clinical prediction modelling. *Family Medicine and Community Health*, v. 8, n. 1, p. e000262, 2020. Disponível em: <https://doi.org/10.1136/fmch-2019-000262>. Acesso em: 5 mar. 2026.

CHRISTODOULOU, E. et al. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *Journal of Clinical Epidemiology*, v. 110, p. 12–22, 2019. Disponível em: <https://doi.org/10.1016/j.jclinepi.2019.02.004>. Acesso em: 5 mar. 2026.

COLAK, A. et al. Diagnostic values of neutrophil/lymphocyte ratio, platelet/lymphocyte ratio and procalcitonin in early diagnosis of bacteremia. *Turkish Journal of Biochemistry*, v. 45, n. 1, p. 57–64, 2020. Disponível em: <https://doi.org/10.1515/tjb-2018-0484>. Acesso em: 5 mar. 2026.

COPAS, J. B. Regression, prediction and shrinkage. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, v. 45, n. 3, p. 311–354, 1983. Disponível em: <https://doi.org/10.1111/j.2517-6161.1983.tb01258.x>. Acesso em: 5 mar. 2026.

CORTES, C.; VAPNIK, V. Support-vector networks. *Machine Learning*, v. 20, p. 273–297, 1995. Disponível em: <https://doi.org/10.1007/BF00994018>. Acesso em: 5 mar. 2026.

COX, D. R. The regression analysis of binary sequences. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, v. 20, n. 2, p. 215–242, 1958. Disponível em: <https://doi.org/10.1111/j.2517-6161.1958.tb00292.x>. Acesso em: 5 mar. 2026.

CUETO-LÓPEZ, N. et al. A comparative study on feature selection for a risk prediction model for colorectal cancer. *Computer Methods and Programs in Biomedicine*, v. 177, p. 219–229, 2019. Disponível em: <https://doi.org/10.1016/j.cmpb.2019.06.001>. Acesso em: 5 mar. 2026.

DAMBROSO-ALTAFINI, D. et al. Routine laboratory biomarkers used to predict Gram-positive or Gram-negative bacteria involved in bloodstream infections. *Scientific Reports*, v. 12, n. 1, p. 15466, 2022. Disponível em: <https://doi.org/10.1038/s41598-022-19643-1>. Acesso em: 5 mar. 2026.

DERKSEN, S.; KESELMAN, H. J. Backward, forward and stepwise automated subset selection algorithms: Frequency of obtaining authentic and noise variables. *British Journal of Mathematical and Statistical Psychology*, v. 45, n. 2, p. 265–282, 1992. Disponível em: <https://doi.org/10.1111/j.2044-8317.1992.tb00992.x>. Acesso em: 5 mar. 2026.

DESBOULETS, L. D. D. A review on variable selection in regression analysis. *Econometrics*, v. 6, n. 4, p. 45, 2018. Disponível em: <https://doi.org/10.3390/econometrics6040045>. Acesso em: 5 mar. 2026.

DIAZ-RAMIREZ, L. G. et al. A novel method for identifying a parsimonious and accurate predictive model for multiple clinical outcomes. *Computer Methods and Programs in Biomedicine*, v. 204, p. 106073, 2021. Disponível em: <https://doi.org/10.1016/j.cmpb.2021.106073>. Acesso em: 5 mar. 2026.

DORMANN, C. F. et al. Collinearity: A review of methods to deal with it and a simulation study evaluating their performance. *Ecography*, v. 36, n. 1, p. 27–46, 2013. Disponível em: <https://doi.org/10.1111/j.1600-0587.2012.07348.x>. Acesso em: 5 mar. 2026.

DUNN, J.; MINGARDI, L.; ZHUO, Y. D. Comparing interpretability and explainability for feature selection. *arXiv preprint arXiv:2105.05328*, 2021. Disponível em: <https://doi.org/10.48550/arXiv.2105.05328>. Acesso em: 5 mar. 2026.

EFRON, B. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, v. 7, n. 1, p. 1–26, 1979. Disponível em: <https://doi.org/10.1214/aos/1176344552>. Acesso em: 5 mar. 2026.

EFRON, B. Computers and the theory of statistics: Thinking the unthinkable. *SIAM Review*, v. 21, n. 4, p. 460–480, 1979. Disponível em: <https://doi.org/10.1137/1021092>. Acesso em: 5 mar. 2026.

EFRON, B.; GONG, G. A leisurely look at the bootstrap, the jackknife, and cross-validation. *The American Statistician*, v. 37, n. 1, p. 36–48, 1983. Disponível em: <https://doi.org/10.2307/2685844>. Acesso em: 5 mar. 2026.

EVANS, L. et al. Surviving sepsis campaign: International guidelines for management of sepsis and septic shock 2021. *Critical Care Medicine*, v. 49, n. 11, p. e1063–e1143, 2021. Disponível em: <https://doi.org/10.1097/CCM.0000000000005337>. Acesso em: 5 mar. 2026.

FISHER, A.; RUDIN, C.; DOMINICI, F. All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research*, v. 20, n. 177, p. 1–81, 2019. Disponível em: <https://doi.org/10.48550/arXiv.1801.01489>. Acesso em: 5 mar. 2026.

FLACK, V. F.; CHANG, P. C. Frequency of selecting noise variables in subset regression analysis: A simulation study. *The American Statistician*, v. 41, n. 1, p. 84–86, 1987. Disponível em: <https://doi.org/10.2307/2684336>. Acesso em: 5 mar. 2026.

FLEUREN, L. M. et al. Machine learning for the prediction of sepsis: A systematic review and meta-analysis of diagnostic test accuracy. *Intensive Care Medicine*, v. 46, n. 3, p. 383–400, 2020. Disponível em: <https://doi.org/10.1007/s00134-019-05872-y>. Acesso em: 5 mar. 2026.

FREUND, Y.; SCHAPIRE, R. E. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, v. 55, n. 1, p. 119–139, 1997. Disponível em: <https://doi.org/10.1006/jcss.1997.1504>. Acesso em: 5 mar. 2026.

FRIEDMAN, J. H. Multivariate adaptive regression splines. *The Annals of Statistics*, v. 19, n. 1, p. 1–141, 1991. Disponível em: <https://doi.org/10.1214/aos/1176347963>. Acesso em: 5 mar. 2026.

FRIEDMAN, J. H. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, v. 29, n. 5, p. 1189–1232, 2001. Disponível em: <https://doi.org/10.1214/aos/1013203451>. Acesso em: 5 mar. 2026.

GAO, Q. et al. Combined procalcitonin and hemogram parameters contribute to early differential diagnosis of Gram-negative/Gram-positive bloodstream infections. *Journal of Clinical Laboratory Analysis*, v. 35, n. 9, p. e23927, 2021. Disponível em: <https://doi.org/10.1002/jcla.23927>. Acesso em: 5 mar. 2026.

GARCÍA-LUQUE, R. et al. A machine learning-based clinical prediction rule for adverse outcomes in multimorbid patients. *Intelligent Medicine*, v. 5, n. 4, p. 300–309, 2025. Disponível em: <https://doi.org/10.1016/j.imed.2025.08.006>. Acesso em: 5 mar. 2026.

GEORGE, E. I. The variable selection problem. *Journal of the American Statistical Association*, v. 95, n. 452, p. 1304–1308, 2000. Disponível em: <https://doi.org/10.2307/2669776>. Acesso em: 5 mar. 2026.

GEURTS, P.; ERNST, D.; WEHENKEL, L. Extremely randomized trees. *Machine Learning*, v. 63, p. 3–42, 2006. Disponível em: <https://doi.org/10.1007/s10994-006-6226-1>. Acesso em: 5 mar. 2026.

GOH, K. H. et al. Artificial intelligence in sepsis early prediction and diagnosis using unstructured data in healthcare. *Nature Communications*, v. 12, n. 1, p. 711, 2021. Disponível em: <https://doi.org/10.1038/s41467-021-20910-4>. Acesso em: 5 mar. 2026.

GONG, G. Some ideas on using the bootstrap in assessing model variability. In: *Computer Science and Statistics: Proceedings of the 14th Symposium on the Interface*. New York: Springer, 1982. p. 169–173.

HANLEY, J. A.; MCNEIL, B. J. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, v. 143, n. 1, p. 29–36, 1982. Disponível em: <https://doi.org/10.1148/radiology.143.1.7063747>. Acesso em: 5 mar. 2026.

HARRELL, F. E. *Regression Modeling Strategies*: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis. 2. ed. Cham: Springer, 2015.

HARTIGAN, J. A. Using subsample values as typical values. *Journal of the American Statistical Association*, v. 64, n. 328, p. 1303–1317, 1969. Disponível em: <http://www.jstor.org/stable/2286069>. Acesso em: 5 mar. 2026.

HARTIGAN, J. A. Necessary and sufficient conditions for asymptotic joint normality of a statistic and its subsample values. *The Annals of Statistics*, v. 3, n. 3, p. 573–580, 1975. Disponível em: <https://doi.org/10.1214/aos/1176343123>. Acesso em: 5 mar. 2026.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. *The Elements of Statistical Learning*: Data Mining, Inference, and Prediction. 2. ed. New York: Springer, 2009.

HEINZE, G.; DUNKLER, D. Five myths about variable selection. *Transplant International*, v. 30, n. 1, p. 6–10, 2017. Disponível em: <https://doi.org/10.1111/tri.12895>. Acesso em: 5 mar. 2026.

HEINZE, G.; WALLISCH, C.; DUNKLER, D. Variable selection—a review and recommendations for the practicing statistician. *Biometrical Journal*, v. 60, n. 3, p. 431–449, 2018. Disponível em: <https://doi.org/10.1002/bimj.201700067>. Acesso em: 5 mar. 2026.

HERTEL, L. et al. Sherpa: Robust hyperparameter optimization for machine learning. *SoftwareX*, v. 12, p. 100591, 2020. Disponível em: <https://doi.org/10.1016/j.softx.2020.100591>. Acesso em: 5 mar. 2026.

HOCKING, R. R. A Biometrics invited paper. The analysis and selection of variables in linear regression. *Biometrics*, p. 1–49, 1976. Disponível em: <https://doi.org/10.2307/2529336>. Acesso em: 5 mar. 2026.

HOERL, A. E.; KENNARD, R. W. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, v. 12, n. 1, p. 55–67, 1970. Disponível em: <https://doi.org/10.2307/1267351>. Acesso em: 5 mar. 2026.

HU, C. et al. Application of machine learning for clinical subphenotype identification in sepsis. *Infectious Diseases and Therapy*, v. 11, n. 5, p. 1949–1964, 2022. Disponível em: <https://doi.org/10.1007/s40121-022-00684-y>. Acesso em: 5 mar. 2026.

HUMMEL, J. et al. Trabecular texture and paraspinal muscle characteristics for prediction of first vertebral fracture: A QCT analysis from the AGES cohort. *Frontiers in Endocrinology*, v. 16, p. 1566424, 2025. Disponível em: <https://doi.org/10.3389/fendo.2025.1566424>. Acesso em: 5 mar. 2026.

HUTTER, F.; KOTTHOFF, L.; VANSCHOREN, J. (Ed.). *Automated Machine Learning*: Methods, Systems, Challenges. Cham: Springer Nature, 2019.

HUTTUNEN, R.; AITTONIEMI, J. New concepts in the pathogenesis, diagnosis and treatment of bacteremia and sepsis. *Journal of Infection*, v. 63, n. 6, p. 407–419, 2011. Disponível em: <https://doi.org/10.1016/j.jinf.2011.08.004>. Acesso em: 5 mar. 2026.

HYDE, R. M. et al. Factors associated with daily weight gain in preweaned calves on dairy farms. *Preventive Veterinary Medicine*, v. 190, p. 105320, 2021. Disponível em: <https://doi.org/10.1016/j.prevetmed.2021.105320>. Acesso em: 5 mar. 2026.

JUUTILAINEN, A. et al. Biomarkers for bacteremia and severe sepsis in hematological patients with neutropenic fever: Multivariate logistic regression analysis and factor analysis. *Leukemia & Lymphoma*, v. 52, n. 12, p. 2349–2355, 2011. Disponível em: <https://doi.org/10.3109/10428194.2011.597904>. Acesso em: 5 mar. 2026.

KAMKAR, I. et al. Stable feature selection for clinical prediction: Exploiting ICD tree structure using Tree-Lasso. *Journal of Biomedical Informatics*, Elsevier, v. 53, p. 277–290, 2015. Disponível em: <https://doi.org/10.1016/j.jbi.2014.11.013>. Acesso em: 10 mar. 2026.

KE, G. et al. LightGBM: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems*, v. 30, p. 3146–3154, 2017. Disponível em: <https://dl.acm.org/doi/10.5555/3294996.3295074>.

KUHN, M. Building predictive models in r using the caret package. *Journal of Statistical Software*, v. 28, n. 5, p. 1–26, 2008. Disponível em: <https://doi.org/10.18637/jss.v028.i05>. Acesso em: 10 mar. 2026.

KWON, Y. et al. Stability selection for LASSO with weights based on AUC. *Scientific Reports*, v. 13, n. 1, p. 5207, 2023. Disponível em: <https://doi.org/10.1038/s41598-023-32517-4>. Acesso em: 5 mar. 2026.

LEE, K. H. et al. Early detection of bacteraemia using ten clinical variables with an artificial neural network approach. *Journal of Clinical Medicine*, v. 8, n. 10, p. 1592, 2019. Disponível em: <https://doi.org/10.3390/jcm8101592>. Acesso em: 5 mar. 2026.

LIEN, F. et al. Bacteremia detection from complete blood count and differential leukocyte count with machine learning: Complementary and competitive with C-reactive protein and procalcitonin tests. *BMC Infectious Diseases*, v. 22, n. 1, p. 287, 2022. Disponível em: <https://doi.org/10.1186/s12879-022-07223-7>. Acesso em: 5 mar. 2026.

LIM, C.; YU, B. Estimation stability with cross-validation (ESCV). *Journal of Computational and Graphical Statistics*, v. 25, n. 2, p. 464–492, 2016. Disponível em: <https://doi.org/10.48550/arXiv.1303.3128>. Acesso em: 5 mar. 2026.

LIN, C.-T. et al. Diagnostic value of serum procalcitonin, lactate, and high-sensitivity C-reactive protein for predicting bacteremia in adult patients in the emergency department. *PeerJ*, v. 5, p. e4094, 2017. Disponível em: <https://doi.org/10.7717/peerj.4094>. Acesso em: 5 mar. 2026.

LIU, X. et al. Clinical validation and optimization of machine learning models for early prediction of sepsis. *Frontiers in Medicine*, v. 12, p. 1521660, 2025. Disponível em: <https://doi.org/10.3389/fmed.2025.1521660>. Acesso em: 5 mar. 2026.

LOONEN, A. J. M. et al. Developments for improved diagnosis of bacterial bloodstream infections. *European Journal of Clinical Microbiology & Infectious Diseases*, v. 33, p. 1687–1702, 2014. Disponível em: <https://doi.org/10.1007/s10096-014-2153-4>. Acesso em: 5 mar. 2026.

MAHMOUD, E. et al. Developing machine-learning prediction algorithm for bacteremia in admitted patients. *Infection and Drug Resistance*, v. 14, p. 757–765, 2021. Disponível em: <https://doi.org/10.2147/IDR.S293496>. Acesso em: 5 mar. 2026.

MAHMOUDIAN, M. et al. Stable iterative variable selection. *Bioinformatics*, v. 37, n. 24, p. 4810–4817, 2021. Disponível em: <https://doi.org/10.1093/bioinformatics/btab501>. Acesso em: 5 mar. 2026.

MANTEL, N. Why stepdown procedures in variable selection. *Technometrics*, v. 12, n. 3, p. 621–625, 1970. Disponível em: <https://doi.org/10.2307/1267207>. Acesso em: 5 mar. 2026.

MATHARAARACHCHI, S.; DOMARATZKI, M.; MUTHUKUMARANA, S. Minimizing features while maintaining performance in data classification problems. *PeerJ Computer Science*, v. 8, p. e1081, 2022. Disponível em: <https://doi.org/10.7717/peerj-cs.1081>. Acesso em: 5 mar. 2026.

MATONO, T. et al. Diagnostic accuracy of quick SOFA score and inflammatory biomarkers for predicting community-onset bacteremia. *Scientific Reports*, v. 12, n. 1, p. 11121, 2022. Disponível em: <https://doi.org/10.1038/s41598-022-15408-y>. Acesso em: 5 mar. 2026.

MCKINNEY, W. et al. pandas: A foundational Python library for data analysis and statistics. In: *Python for High Performance and Scientific Computing*. [S.l.: s.n.], 2011. v. 14, n. 9.

MEINSHAUSEN, N.; BÜHLMANN, P. Stability selection. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, v. 72, n. 4, p. 417–473, 2010. Disponível em: <https://doi.org/https://doi.org/10.48550/arXiv.0809.2932>. Acesso em: 5 mar. 2026.

MILLER, A. *Subset Selection in Regression*. 2. ed. New York: Chapman and Hall/CRC, 2002.

MOOR, M. et al. Early prediction of sepsis in the ICU using machine learning: A systematic review. *Frontiers in Medicine*, v. 8, p. 607952, 2021. Disponível em: <https://doi.org/10.3389/fmed.2021.607952>. Acesso em: 5 mar. 2026.

MURRI, R. et al. A machine learning predictive model of bloodstream infection in hospitalized patients. *Diagnostics*, v. 14, n. 4, p. 445, 2024. Disponível em: <https://doi.org/10.3390/diagnostics14040445>. Acesso em: 5 mar. 2026.

MURTAUGH, P. A. Methods of variable selection in regression modeling. *Communications in Statistics-Simulation and Computation*, v. 27, n. 3, p. 711–734, 1998. Disponível em: <https://doi.org/10.1080/03610919808813505>. Acesso em: 5 mar. 2026.

NEŠKOVIĆ, N. et al. Predictive role of selected biomarkers in differentiating Gram-positive from Gram-negative sepsis in surgical patients: A retrospective study. *Anaesthesiology Intensive Therapy*, v. 55, n. 5, p. 319–325, 2023. Disponível em: <https://doi.org/10.5114/ait.2023.134214>. Acesso em: 5 mar. 2026.

NISHIKAWA, H. et al. Comparative usefulness of inflammatory markers to indicate bacterial infection—analyzed according to blood culture results and related clinical factors. *Diagnostic Microbiology and Infectious Disease*, v. 84, n. 1, p. 69–73, 2016. Disponível em: <https://doi.org/10.1016/j.diagmicrobio.2015.09.015>. Acesso em: 5 mar. 2026.

NOGUEIRA, S.; SECHIDIS, K.; BROWN, G. On the stability of feature selection algorithms. *Journal of Machine Learning Research*, v. 18, n. 174, p. 1–54, 2018. Disponível em: <https://dl.acm.org/doi/10.5555/3122009.3242031>. Acesso em: 5 mar. 2026.

O'REILLY, D.; MCGRATH, J.; MARTIN-LOECHES, I. Optimizing artificial intelligence in sepsis management: Opportunities in the present and looking closely to the future. *Journal of Intensive Medicine*, v. 4, n. 1, p. 34–45, 2024. Disponível em: <https://doi.org/10.1016/j.jointm.2023.10.001>. Acesso em: 5 mar. 2026.

ORLENKO, A.; MOORE, J. H. A comparison of methods for interpreting random forest models of genetic association in the presence of non-additive interactions. *BioData Mining*, v. 14, n. 1, p. 9, 2021. Disponível em: <https://doi.org/10.1186/s13040-021-00243-0>. Acesso em: 5 mar. 2026.

OTTARSDOTTIR, K. et al. Cardiometabolic risk factors and endogenous sex hormones in postmenopausal women: A cross-sectional study. *Journal of the Endocrine Society*, v. 6, n. 6, p. bvac050, 2022. Disponível em: <https://doi.org/10.1210/jendso/bvac050>. Acesso em: 5 mar. 2026.

PAIVA, P. P. et al. Risk-prediction model for transfusion of erythrocyte concentrate during extracorporeal circulation in coronary surgery. *Brazilian Journal of Cardiovascular Surgery*, v. 36, n. 3, p. 323–330, 2021. Disponível em: <https://doi.org/10.21470/1678-9741-2020-0322>. Acesso em: 5 mar. 2026.

PEDREGOSA, F. et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, v. 12, p. 2825–2830, 2011. Disponível em: <https://doi.org/10.48550/arXiv.1201.0490>. Acesso em: 5 mar. 2026.

PETERS, R. P. H. et al. New developments in the diagnosis of bloodstream infections. *The Lancet Infectious Diseases*, v. 4, n. 12, p. 751–760, 2004. Disponível em: <https://doi.org/10.1016/S1473-3099(04)01205-8>. Acesso em: 5 mar. 2026.

POLITIS, D. N.; ROMANO, J. P. Large sample confidence regions based on subsamples under minimal assumptions. *The Annals of Statistics*, v. 22, n. 4, p. 2031–2050, 1994. Disponível em: <https://doi.org/10.1214/aos/1176325770>. Acesso em: 5 mar. 2026.

POLITIS, D. N.; ROMANO, J. P.; WOLF, M. *Subsampling*. New York: Springer, 1999.

PRESCOTT, L. M.; HARLEY, J. P.; KLEIN, D. A. *Microbiology*. New York: McGraw-Hill, 2002.

Python Software Foundation. *Python: A dynamic, open source programming language*. 2023. Disponível em: <https://www.python.org>.

R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2024. Disponível em: <https://www.R-project.org/>.

RATZINGER, F. et al. A risk prediction model for screening bacteremic patients: A cross sectional study. *PLOS ONE*, v. 9, n. 9, p. e106765, 2014. Disponível em: <https://doi.org/10.1371/journal.pone.0106765>. Acesso em: 5 mar. 2026.

RATZINGER, F. et al. Neither single nor a combination of routine laboratory parameters can discriminate between Gram-positive and Gram-negative bacteremia. *Scientific Reports*, v. 5, n. 1, p. 16008, 2015. Disponível em: <https://doi.org/10.1038/srep16008>. Acesso em: 5 mar. 2026.

RILEY, R. D.; COLLINS, G. S. Stability of clinical prediction models developed using statistical or machine learning methods. *Biometrical Journal*, v. 65, n. 8, p. 2200302, 2023. Disponível em: <https://doi.org/10.1002/bimj.202200302>. Acesso em: 5 mar. 2026.

ROBIN, X. et al. proc: an open-source package for r and s+ to analyze and compare roc curves. *BMC Bioinformatics*, v. 12, p. 77, 2011. Disponível em: <https://doi.org/10.1186/1471-2105-12-77>.

SAUERBREI, W.; BOULESTEIX, A.-L.; BINDER, H. Stability investigations of multivariable regression models derived from low- and high-dimensional data. *Journal of Biopharmaceutical Statistics*, v. 21, n. 6, p. 1206–1231, 2011. Disponível em: <https://doi.org/10.1080/10543406.2011.629890>. Acesso em: 5 mar. 2026.

SAUERBREI, W. et al. On stability issues in deriving multivariable regression models. *Biometrical Journal*, v. 57, n. 4, p. 531–555, 2015. Disponível em: <https://doi.org/10.1002/bimj.201300222>. Acesso em: 5 mar. 2026.

SAUERBREI, W.; ROYSTON, P. Building multivariable prognostic and diagnostic models: Transformation of the predictors by using fractional polynomials. *Journal of the Royal Statistical Society Series A: Statistics in Society*, v. 162, n. 1, p. 71–94, 1999. Disponível em: <https://doi.org/10.1111/1467-985X.00122>. Acesso em: 5 mar. 2026.

SAUERBREI, W.; SCHUMACHER, M. A bootstrap resampling procedure for model building: Application to the Cox regression model. *Statistics in Medicine*, v. 11, n. 16, p. 2093–2109, 1992. Disponível em: <https://doi.org/10.1002/sim.4780111607>. Acesso em: 5 mar. 2026.

SCHINKEL, M. et al. Diagnostic stewardship for blood cultures in the emergency department: A multicenter validation and prospective evaluation of a machine learning prediction tool. *EBioMedicine*, v. 82, 2022. Disponível em: <https://doi.org/10.1016/j.ebiom.2022.104176>. Acesso em: 5 mar. 2026.

SCHWARZ, G. Estimating the dimension of a model. *The Annals of Statistics*, p. 461–464, 1978. Disponível em: <https://doi.org/10.1214/aos/1176344136>. Acesso em: 5 mar. 2026.

SHAO, H. et al. Optimization of diabetes prediction methods based on combinatorial balancing algorithm. *Nutrition & Diabetes*, v. 14, n. 1, p. 63, 2024. Disponível em: <https://doi.org/10.1038/s41387-024-00324-z>. Acesso em: 5 mar. 2026.

SHIM, B.-S. et al. Clinical value of whole blood procalcitonin using point of care testing, quick sequential organ failure assessment score, C-reactive protein and lactate in emergency department patients with suspected infection. *Journal of Clinical Medicine*, v. 8, n. 6, p. 833, 2019. Disponível em: <https://doi.org/10.3390/jcm8060833>. Acesso em: 5 mar. 2026.

SINGER, M. et al. The third international consensus definitions for sepsis and septic shock (Sepsis-3). *JAMA*, v. 315, n. 8, p. 801–810, 2016. Disponível em: <https://doi.org/10.1001/jama.2016.0287>. Acesso em: 5 mar. 2026.

SONG, X. et al. Comparison of machine learning and logistic regression models in predicting acute kidney injury: A systematic review and meta-analysis. *International Journal of Medical Informatics*, v. 151, p. 104484, 2021. Disponível em: <https://doi.org/10.1016/j.ijmedinf.2021.104484>. Acesso em: 5 mar. 2026.

SRISKANDAN, S.; COHEN, J. Gram-positive sepsis: Mechanisms and differences from Gram-negative sepsis. *Infectious Disease Clinics of North America*, v. 13, n. 2, p. 397–412, 1999. Disponível em: <https://doi.org/10.1016/s0891-5520(05)70082-9>. Acesso em: 5 mar. 2026.

STEEN, K. V. et al. Multicollinearity in prognostic factor analyses using the EORTC QLQ-C30: Identification and impact on model selection. *Statistics in Medicine*, v. 21, n. 24, p. 3865–3884, 2002. Disponível em: <https://doi.org/10.1002/sim.1358>. Acesso em: 5 mar. 2026.

STEENKISTE, T. V. et al. Accurate prediction of blood culture outcome in the intensive care unit using long short-term memory neural networks. *Artificial Intelligence in Medicine*, v. 97, p. 38–43, 2019. Disponível em: <https://doi.org/10.1016/j.artmed.2018.10.008>. Acesso em: 5 mar. 2026.

STEYERBERG, E. W. *Clinical Prediction Models*: A Practical Approach to Development, Validation, and Updating. 2. ed. Cham: Springer, 2019.

STEYERBERG, E. W. et al. Prognostic modeling with logistic regression analysis: In search of a sensible strategy in small data sets. *Medical Decision Making*, v. 21, n. 1, p. 45–56, 2001. Disponível em: <https://doi.org/10.1177/0272989X0102100106>. Acesso em: 5 mar. 2026.

STEYERBERG, E. W. et al. Assessing the performance of prediction models: A framework for traditional and novel measures. *Epidemiology*, v. 21, n. 1, p. 128–138, 2010. Disponível em: <https://doi.org/10.1097/EDE.0b013e3181c30fb2>. Acesso em: 5 mar. 2026.

SU, M. et al. Four biomarkers-based artificial neural network model for accurate early prediction of bacteremia with low-level procalcitonin. *Annals of Clinical & Laboratory Science*, v. 51, n. 3, p. 408–414, 2021.

TANG, A. et al. Prognostic differences in sepsis caused by Gram-negative bacteria and Gram-positive bacteria: A systematic review and meta-analysis. *Critical Care*, v. 27, n. 1, p. 467, 2023. Disponível em: <https://doi.org/10.1186/s13054-023-04750-w>. Acesso em: 5 mar. 2026.

TANG, W. et al. Hematological parameters in patients with bloodstream infection: A retrospective observational study. *The Journal of Infection in Developing Countries*, v. 14, n. 11, p. 1264–1273, 2020. Disponível em: <https://doi.org/10.3855/jidc.12811>. Acesso em: 5 mar. 2026.

THENG, D.; BHOYAR, K. K. Feature selection techniques for machine learning: A survey of more than two decades of research. *Knowledge and Information Systems*, v. 66, n. 3, p. 1575–1637, 2024. Disponível em: <https://doi.org/10.1007/s10115-023-02010-5>. Acesso em: 5 mar. 2026.

THOMAS-RÜDDEL, D. O. et al. Influence of pathogen and focus of infection on procalcitonin values in sepsis patients with bacteremia or candidemia. *Critical Care*, v. 22, n. 1, p. 128, 2018. Disponível em: <https://doi.org/10.1186/s13054-018-2050-9>. Acesso em: 5 mar. 2026.

TIBSHIRANI, R. Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, v. 58, n. 1, p. 267–288, 1996. Disponível em: <https://doi.org/10.1111/j.2517-6161.1996.tb02080.x>. Acesso em: 5 mar. 2026.

TUKEY, J. W. Some thoughts on clinical trials, especially problems of multiplicity. *Science*, v. 198, n. 4318, p. 679–684, 1977. Disponível em: <https://doi.org/10.1126/science.333584>. Acesso em: 5 mar. 2026.

UYAR, N. Y. et al. Sepsis biomarkers for early diagnosis of bacteremia in emergency department. *The Journal of Infection in Developing Countries*, v. 17, n. 6, p. 832–839, 2023. Disponível em: <https://doi.org/10.3855/jidc.17221>. Acesso em: 5 mar. 2026.

WALLISCH, C. et al. Selection of variables for multivariable models: Opportunities and limitations in quantifying model stability by resampling. *Statistics in Medicine*, v. 40, n. 2, p. 369–381, 2021. Disponível em: <https://doi.org/10.1002/sim.8779>. Acesso em: 5 mar. 2026.

WANG, K. et al. Which biomarkers reveal neonatal sepsis? *PLOS ONE*, v. 8, n. 12, p. e82700, 2013. Disponível em: <https://doi.org/10.1371/journal.pone.0082700>. Acesso em: 5 mar. 2026.

WANG, T.-Q. et al. Prediction of bacteremia using routine hematological and metabolic parameters based on logistic regression and random forest models. *Frontiers in Cellular and Infection Microbiology*, v. 15, p. 1605485, 2025. Disponível em: <https://doi.org/10.3389/fcimb.2025.1605485>. Acesso em: 5 mar. 2026.

WILLEY, J. M.; SANDMAN, K. M.; WOOD, D. H. *Prescott's Microbiology*. 11. ed. New York: McGraw-Hill Education, 2020.

YADGAROV, M. Y. et al. Early detection of sepsis using machine learning algorithms: A systematic review and network meta-analysis. *Frontiers in Medicine*, v. 11, p. 1491358, 2024. Disponível em: <https://doi.org/10.3389/fmed.2024.1491358>. Acesso em: 5 mar. 2026.

YANG, C.-Y. et al. Differential effects of inappropriate empirical antibiotic therapy in adults with community-onset Gram-positive and Gram-negative aerobe bacteremia. *Journal of Infection and Chemotherapy*, v. 26, n. 2, p. 222–229, 2020. Disponível em: <https://doi.org/10.1016/j.jiac.2019.08.021>. Acesso em: 5 mar. 2026.

YIN, Q.-Y.; LI, J.-L.; ZHANG, C.-X. Ensembling variable selectors by stability selection for the Cox model. *Computational Intelligence and Neuroscience*, v. 2017, n. 1, p. 2747431, 2017. Disponível em: <https://doi.org/10.1155/2017/2747431>. Acesso em: 5 mar. 2026.

YU, B. Stability. *Bernoulli*, v. 19, n. 4, p. 1484–1500, 2013. Disponível em: <https://doi.org/10.3150/13-BEJSP14>. Acesso em: 5 mar. 2026.

YU, S. et al. Accurate breast cancer diagnosis using a stable feature ranking algorithm. *BMC Medical Informatics and Decision Making*, v. 23, n. 1, p. 64, 2023. Disponível em: <https://doi.org/10.1186/s12911-023-02142-2>. Acesso em: 5 mar. 2026.

ZHANG, F. et al. Machine learning model for the prediction of Gram-positive and Gram-negative bacterial bloodstream infection based on routine laboratory parameters. *BMC Infectious Diseases*, v. 23, n. 1, p. 675, 2023. Disponível em: <https://doi.org/10.1186/s12879-023-08602-4>. Acesso em: 5 mar. 2026.

ZHANG, S.-Y. et al. Identifying key blood markers for bacteremia in elderly patients: Insights into bacterial pathogens. *Frontiers in Cellular and Infection Microbiology*, v. 14, p. 1472765, 2025. Disponível em: <https://doi.org/10.3389/fcimb.2024.1472765>. Acesso em: 5 mar. 2026.

ZHOU, T. et al. Early identification of bloodstream infection in hemodialysis patients by machine learning. *Heliyon*, v. 9, n. 7, p. e18263, 2023. Disponível em: <https://doi.org/10.1016/j.heliyon.2023.e18263>. Acesso em: 5 mar. 2026.

ZOU, H.; HASTIE, T. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, v. 67, n. 2, p. 301–320, 2005. Disponível em: <https://doi.org/10.1111/j.1467-9868.2005.00503.x>. Acesso em: 5 mar. 2026.

Apêndices

APÊNDICE A

Análise descritiva dos dados

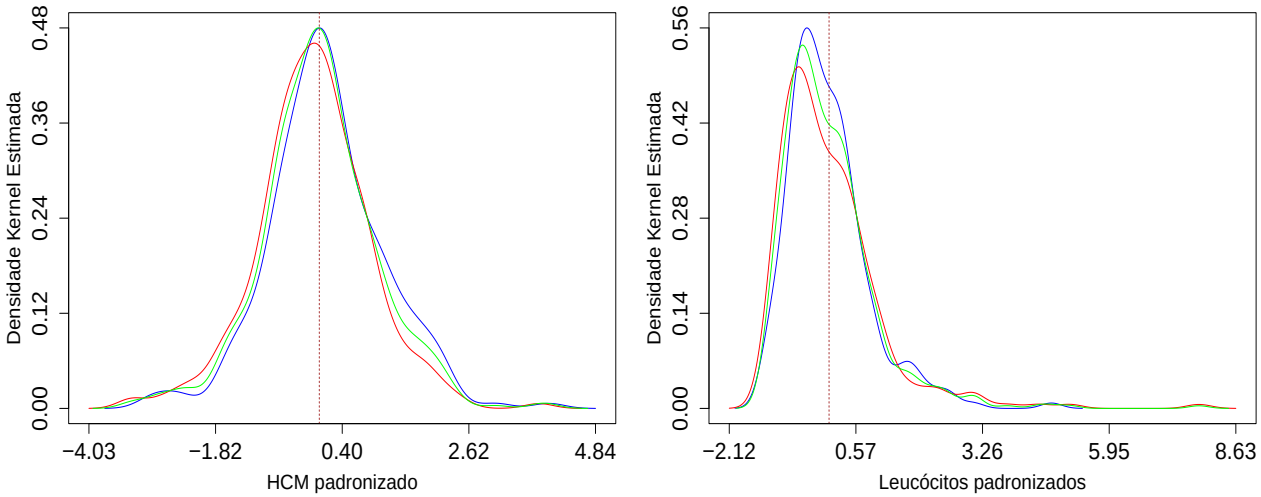

Figura 8 – Densidade Kernel estimada para as variáveis padronizadas HCM e Leucócitos (gram-negativo, gram-positivo, global, média).

Tabela 4 – Descrição das variáveis HCM e Leucócitos.

	HCM		Leucócitos	
Medidas	Gram-negativo	Gram-positivo	Gram-negativo	Gram-positivo
Válidos	261	229	261	229
%Válidos	94.224	93.089	94.224	93.089
Perdidos	16	17	16	17
%Perdidos	5.776	6.911	5.776	6.911
%Zeros	0.000	0.000	0.000	0.000
%(Z-P)	5.776	6.911	5.776	6.911
Média	29.127	29.776	14610.605	14417.817
D.P.	2.855	2.930	11868.654	9209.930
Mínimo	19.900	20.900	250.000	210.000
P10%	25.900	26.560	3490.000	5112.000
P25%	27.500	28.000	6760.000	7800.000
P50%	29.100	29.500	12310.000	12660.000
P75%	30.800	31.500	19430.000	18170.000
P90%	32.600	33.720	25960.000	24942.000
Máximo	40.600	41.100	98490.000	65000.000
ASC-ROC	0.56311		0.52193	
Distância L^2	0.00141		0.00250	

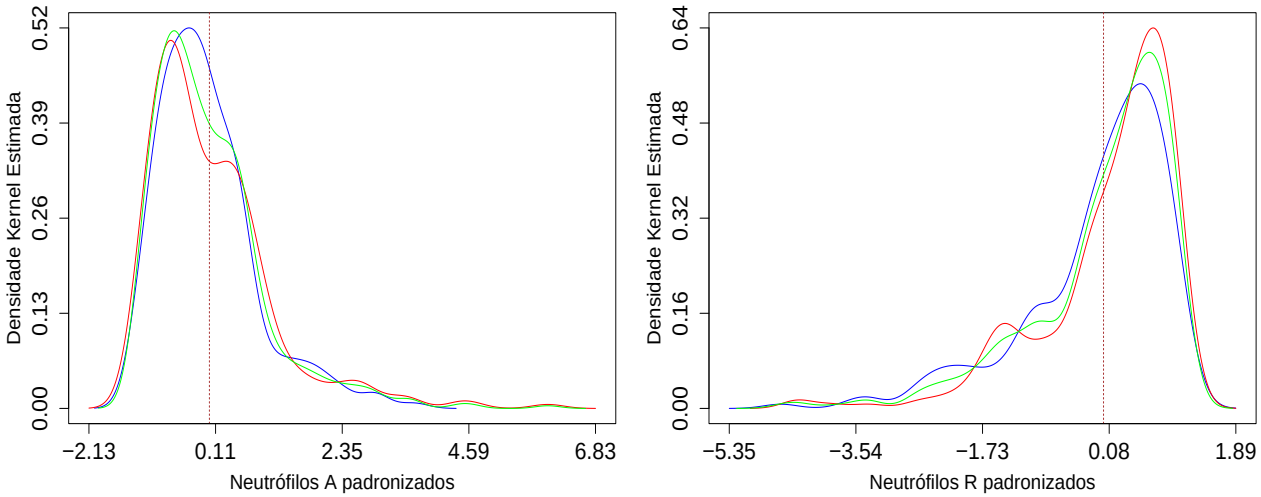

Figura 9 – Densidade Kernel estimada para as variáveis padronizadas Neutrófilos A e Neutrófilos R (gram-negativo, gram-positivo, global, média).

Tabela 5 – Descrição das variáveis Neutrófilos A e Neutrófilos R.

	Neutrófilos A		Neutrófilos R	
Medidas	Gram-negativo	Gram-positivo	Gram-negativo	Gram-positivo
Válidos	261	229	261	229
%Válidos	94.224	93.089	94.224	93.089
Perdidos	16	17	16	17
%Perdidos	5.776	6.911	5.776	6.911
%Zeros	0.000	0.000	0.000	0.000
%(Z-P)	5.776	6.911	5.776	6.911
Média	11988.716	11347.930	79.096	76.301
D.P.	9924.085	7920.401	16.214	17.002
Mínimo	71.000	4.000	5.000	1.000
P10%	2518.000	2992.200	55.000	52.000
P25%	5057.000	5553.000	74.000	70.000
P50%	9444.000	10344.000	84.000	81.000
P75%	16582.000	15388.000	91.000	88.000
P90%	22066.000	21397.800	94.000	93.000
Máximo	65768.000	44443.000	98.000	97.000
ASC-ROC	0.50197		0.56048	
Distância L^2	0.00199		0.00184	

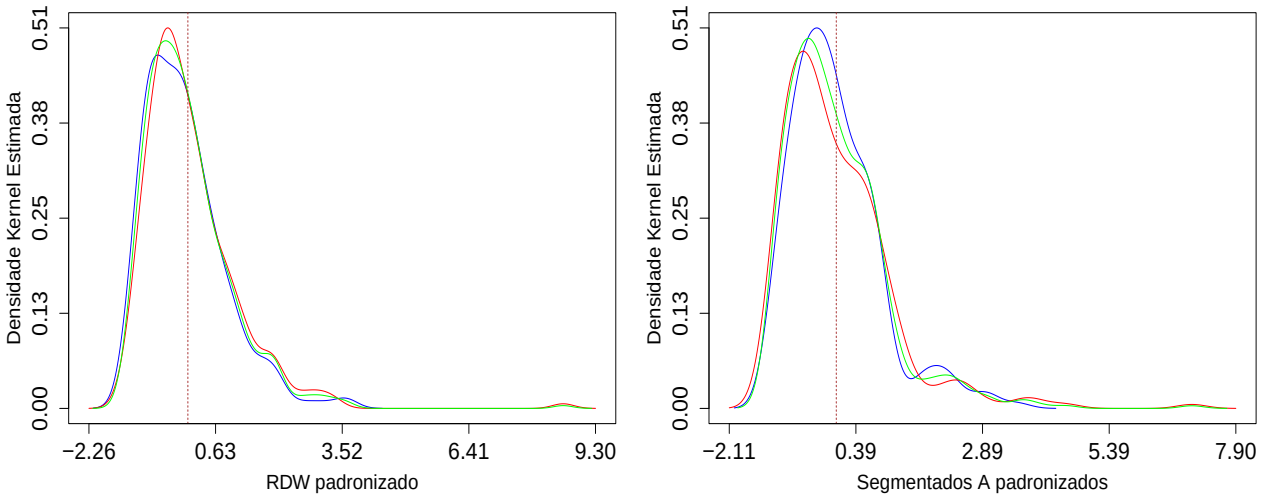

Figura 10 – Densidade Kernel estimada para as variáveis padronizadas RDW e Segmentados A (gram-negativo, gram-positivo, global, média).

Tabela 6 – Descrição das variáveis RDW e Segmentados A.

	RDW		Segmentados A	
Medidas	Gram-negativo	Gram-positivo	Gram-negativo	Gram-positivo
Válidos	261	229	261	229
%Válidos	94.224	93.089	94.224	93.089
Perdidos	16	17	16	17
%Perdidos	5.776	6.911	5.776	6.911
%Zeros	0.000	0.000	0.000	0.000
%(Z-P)	5.776	6.911	5.776	6.911
Média	16.254	15.887	9822.192	9679.555
D.P.	2.879	2.444	8470.064	6915.394
Mínimo	12.000	12.400	71.000	4.500
P10%	13.400	13.300	1404.900	2313.120
P25%	14.400	14.000	3698.400	4665.600
P50%	15.600	15.500	7835.200	8342.400
P75%	17.400	17.100	13899.600	13372.800
P90%	19.500	19.000	18738.000	16976.860
Máximo	39.100	25.900	64383.900	37566.900
ASC-ROC	0.53557		0.51483	
Distância L^2	0.00146		0.00185	

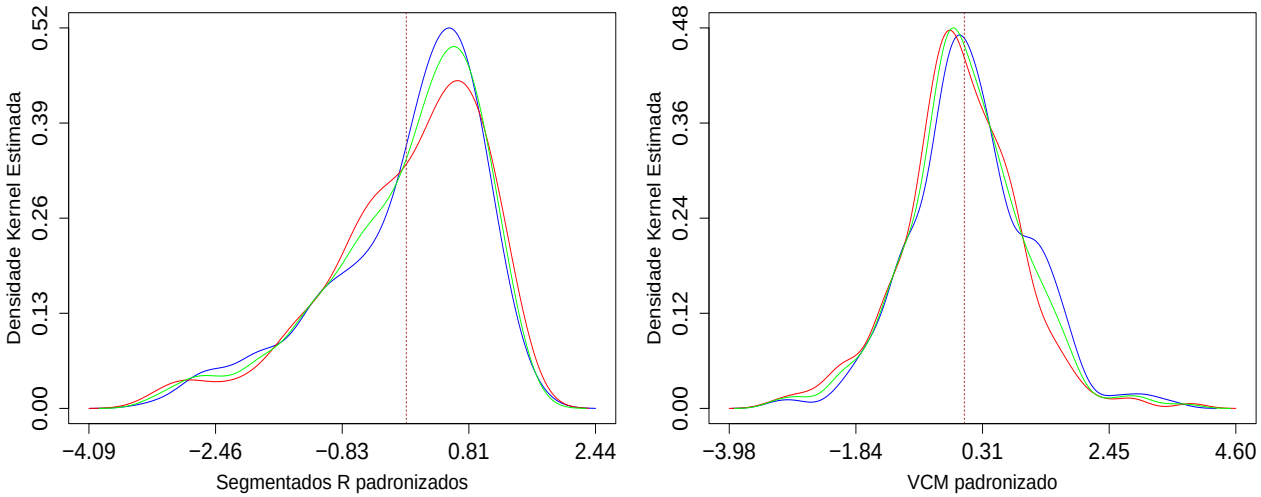

Figura 11 – Densidade Kernel estimada para as variáveis padronizadas Segmentados R e VCM (gram-negativo, gram-positivo, global, média).

Tabela 7 – Descrição das variáveis Segmentados R e VCM.

	Segmentados R		VCM	
Medidas	Gram-negativo	Gram-positivo	Gram-negativo	Gram-positivo
Válidos	261	229	261	229
%Válidos	94.224	93.089	94.224	93.089
Perdidos	16	17	16	17
%Perdidos	5.776	6.911	5.776	6.911
%Zeros	0.000	0.000	0.000	0.000
%(Z-P)	5.776	6.911	5.776	6.911
Média	64.812	64.624	87.844	89.258
D.P.	20.016	19.563	7.965	8.159
Mínimo	1.000	1.000	62.500	62.900
P10%	37.000	37.800	78.600	79.680
P25%	54.000	54.000	83.500	84.800
P50%	70.000	70.000	87.600	88.600
P75%	80.000	79.000	92.800	93.900
P90%	87.000	84.200	96.600	99.540
Máximo	94.000	96.000	119.500	116.800
ASC-ROC	0.49445		0.54764	
Distância L^2	0.00108		0.00144	

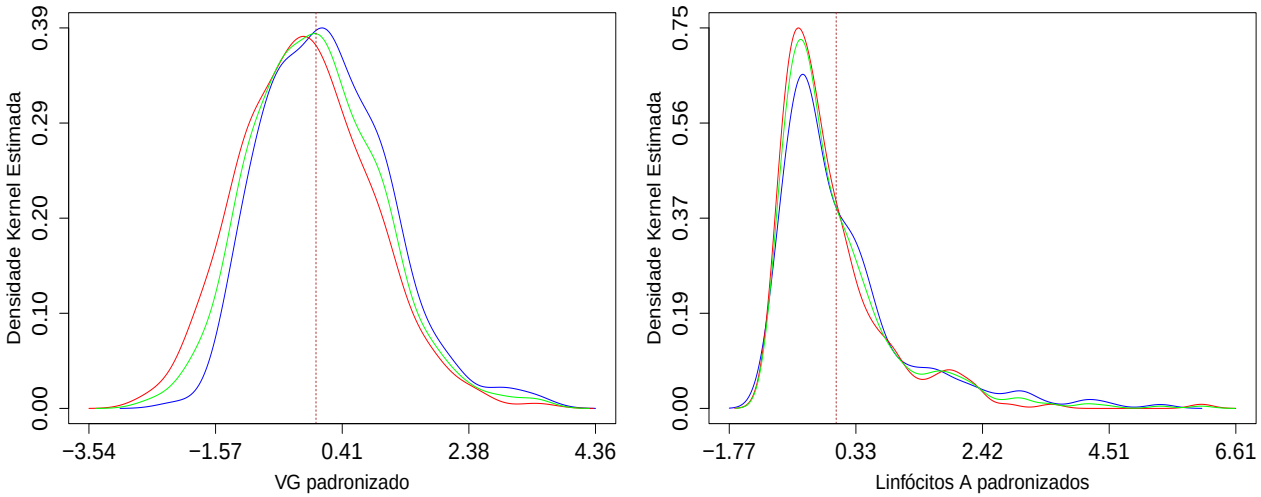

Figura 12 – Densidade Kernel estimada para as variáveis padronizadas VG e Linfócitos A (gram-negativo, gram-positivo, global, média).

Tabela 8 – Descrição das variáveis VG e Linfócitos A.

	VG		Linfócitos A	
Medidas	Gram-negativo	Gram-positivo	Gram-negativo	Gram-positivo
Válidos	261	229	261	229
%Válidos	94.224	93.089	94.224	93.089
Perdidos	16	17	16	17
%Perdidos	5.776	6.911	5.776	6.911
%Zeros	0.000	0.000	0.361	0.000
%(Z-P)	5.776	6.911	6.137	6.911
Média	30.392	32.717	1448.298	1708.793
D.P.	7.065	6.768	1282.046	1547.034
Mínimo	12.900	16.200	0.000	39.900
P10%	21.800	24.500	354.600	383.360
P25%	25.200	27.700	605.000	715.000
P50%	30.300	32.200	1050.500	1204.000
P75%	34.800	37.300	1824.200	2147.200
P90%	39.500	40.760	3049.500	3836.320
Máximo	55.600	55.900	10119.900	9158.400
ASC-ROC	0.59141		0.54764	
Distância L^2	0.00177		0.00170	

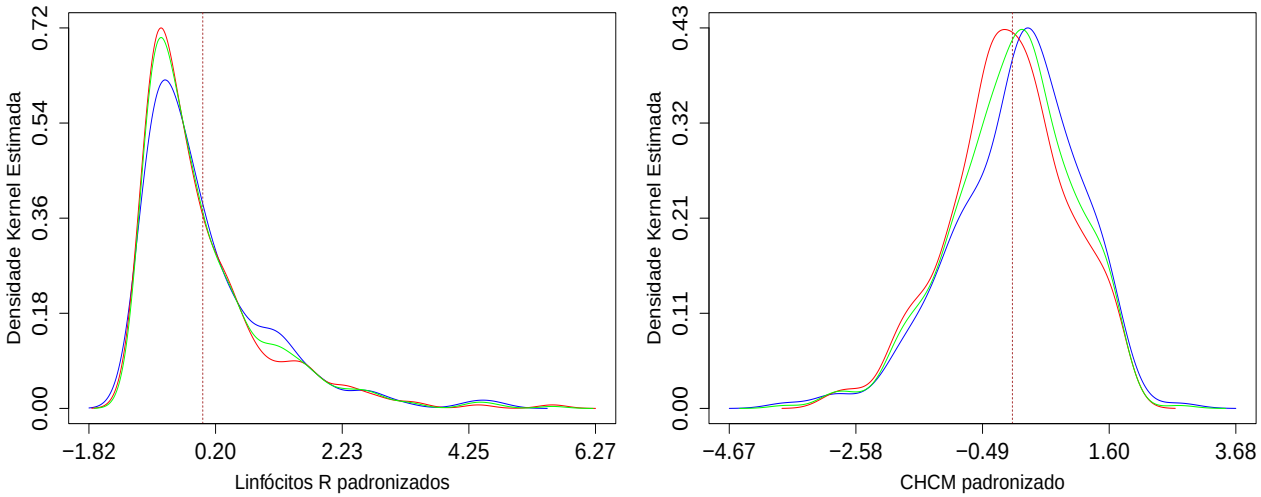

Figura 13 – Densidade Kernel estimada para as variáveis padronizadas Linfócitos R e CHCM (gram-negativo, gram-positivo, global, média).

Tabela 9 – Descrição das variáveis Linfócitos R e CHCM.

	Linfócitos R		CHCM	
Medidas	Gram-negativo	Gram-positivo	Gram-negativo	Gram-positivo
Válidos	261	229	261	228
%Válidos	94.224	93.089	94.224	92.683
Perdidos	16	17	16	18
%Perdidos	5.776	6.911	5.776	7.317
%Zeros	0.361	0.000	0.000	0.000
%(Z-P)	6.137	6.911	5.776	7.317
Média	14.176	15.197	33.149	33.382
D.P.	13.346	13.680	1.504	1.572
Mínimo	0.000	1.000	28.600	27.500
P10%	3.000	3.000	31.000	31.270
P25%	5.000	5.000	32.300	32.400
P50%	10.000	10.000	33.200	33.600
P75%	19.000	21.000	34.100	34.500
P90%	33.000	32.200	35.200	35.400
Máximo	90.000	78.000	36.200	37.500
ASC-ROC	0.52601		0.55249	
Distância L^2	0.00126		0.00180	

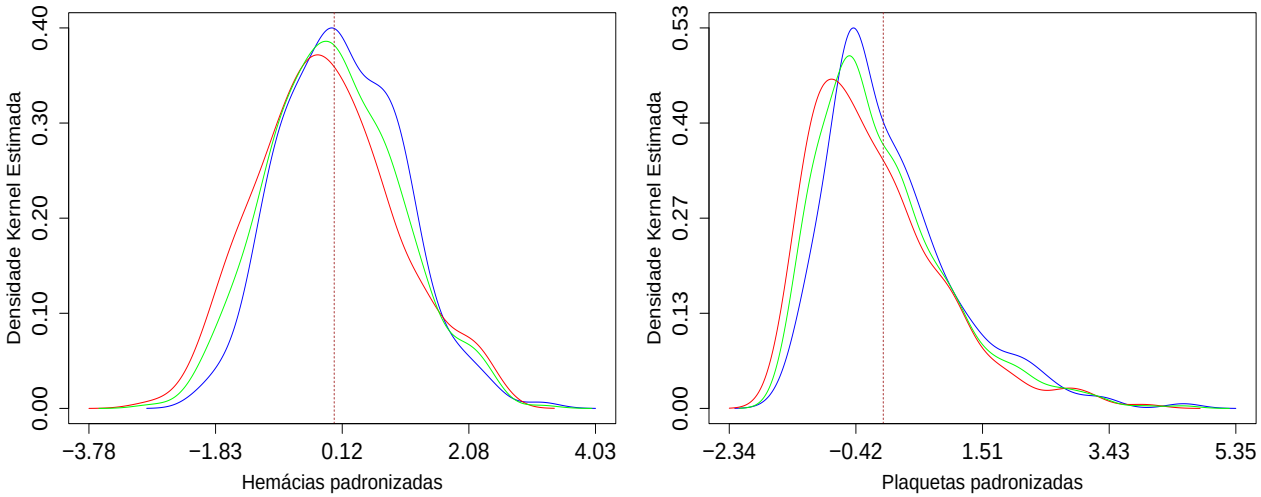

Figura 14 – Densidade Kernel estimada para as variáveis padronizadas Hemácias e Plaquetas (gram-negativo, gram-positivo, global, média).

Tabela 10 – Descrição das variáveis Hemácias e Plaquetas.

	Hemácias		Plaquetas	
Medidas	Gram-negativo	Gram-positivo	Gram-negativo	Gram-positivo
Válidos	261	228	261	228
%Válidos	94.224	92.683	94.224	92.683
Perdidos	16	18	16	18
%Perdidos	5.776	7.317	5.776	7.317
%Zeros	0.000	0.000	0.000	0.000
%(Z-P)	5.776	7.317	5.776	7.317
Média	3.487	3.678	201049.808	234285.088
D.P.	0.850	0.741	141926.286	141188.200
Mínimo	1.280	1.920	5000.000	9000.000
P10%	2.380	2.760	50000.000	85900.000
P25%	2.900	3.140	90000.000	138750.000
P50%	3.410	3.625	169000.000	203000.000
P75%	4.050	4.203	273000.000	305250.000
P90%	4.660	4.583	378000.000	421300.000
Máximo	5.560	6.150	780000.000	865000.000
ASC-ROC	0.57082		0.58083	
Distância L^2	0.00193		0.00240	

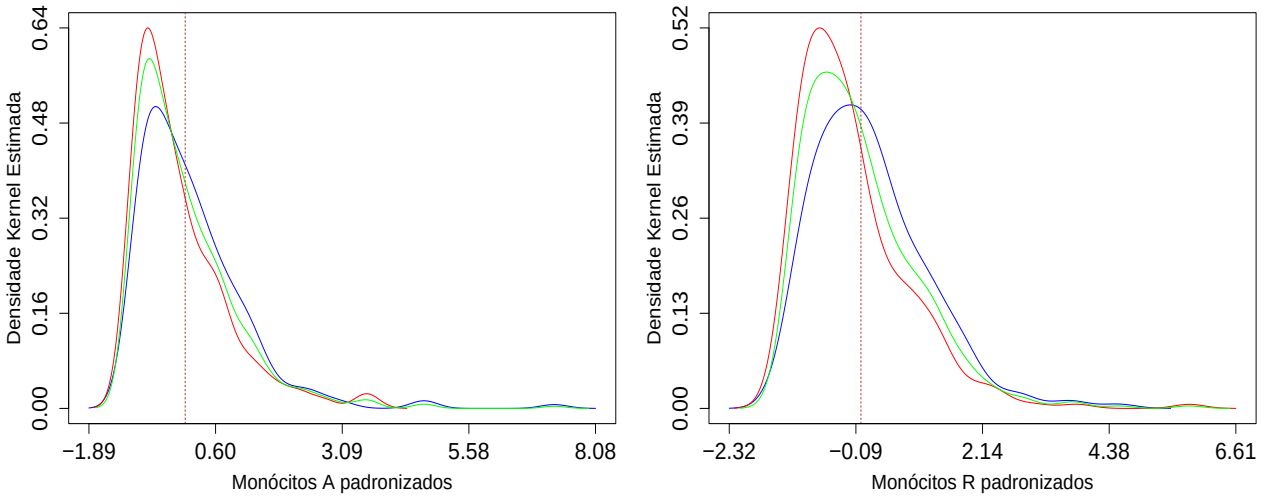

Figura 15 – Densidade Kernel estimada para as variáveis padronizadas Monócitos A e Monócitos R (gram-negativo, gram-positivo, global, média).

Tabela 11 – Descrição das variáveis Monócitos A e Monócitos R.

	Monócitos A		Monócitos R	
Medidas	Gram-negativo	Gram-positivo	Gram-negativo	Gram-positivo
Válidos	261	229	261	229
%Válidos	94.224	93.089	94.224	93.089
Perdidos	16	17	16	17
%Perdidos	5.776	6.911	5.776	6.911
%Zeros	1.805	1.220	1.805	1.220
%(Z-P)	7.581	8.130	7.581	8.130
Média	699.337	864.379	5.057	6.380
D.P.	665.671	767.063	3.922	4.040
Mínimo	0.000	0.000	0.000	0.000
P10%	66.400	148.140	1.000	2.000
P25%	223.800	339.600	2.000	3.000
P50%	508.800	712.500	4.000	6.000
P75%	1009.800	1200.600	7.000	9.000
P90%	1558.400	1689.120	10.000	12.000
Máximo	3393.500	5997.600	29.000	24.000
ASC-ROC	0.57982		0.61150	
Distância L^2	0.00291		0.00331	

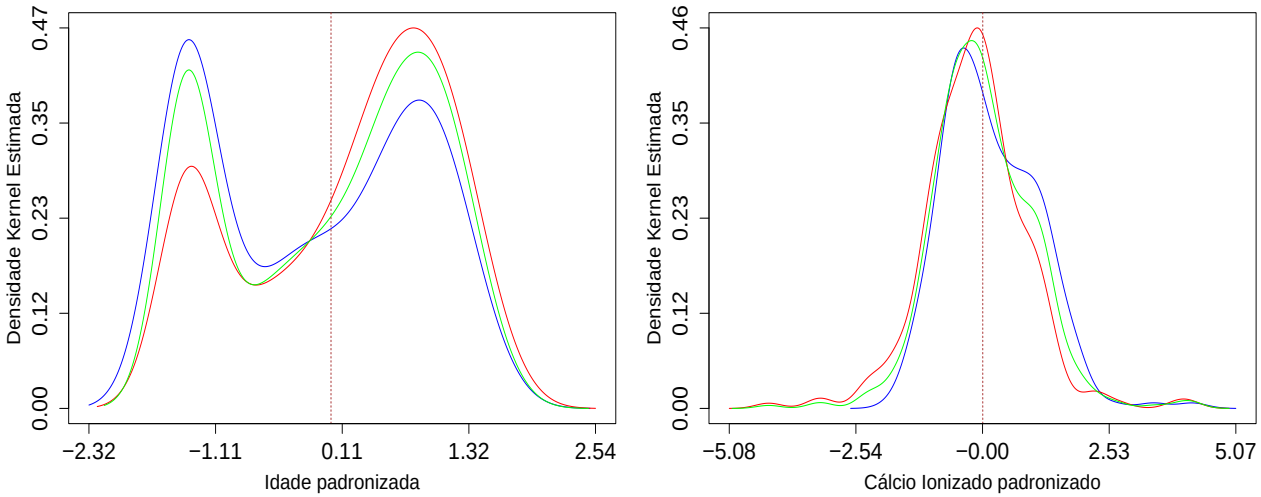

Figura 16 – Densidade Kernel estimada para as variáveis padronizadas Idade e Cálcio Ionizado (gram-negativo, gram-positivo, global, média).

Tabela 12 – Descrição das variáveis Idade e Cálcio Ionizado.

	Idade		Cálcio Ionizado	
Medidas	Gram-negativo	Gram-positivo	Gram-negativo	Gram-positivo
Válidos	277	246	238	211
%Válidos	100.000	100.000	85.921	85.772
Perdidos	0	0	39	35
%Perdidos	0.000	0.000	14.079	14.228
%Zeros	4.693	18.699	0.000	0.000
%(Z-P)	4.693	18.699	14.079	14.228
Média	47.069	37.424	4.646	4.805
D.P.	29.010	31.013	0.511	0.463
Mínimo	0.000	0.000	2.600	3.840
P10%	0.360	0.000	4.150	4.280
P25%	23.000	0.300	4.340	4.445
P50%	56.000	40.000	4.650	4.730
P75%	70.000	66.750	4.920	5.140
P90%	81.000	75.500	5.233	5.390
Máximo	94.000	90.000	6.770	6.800
ASC-ROC	0.59624		0.58481	
Distância L^2	0.00156		0.00239	

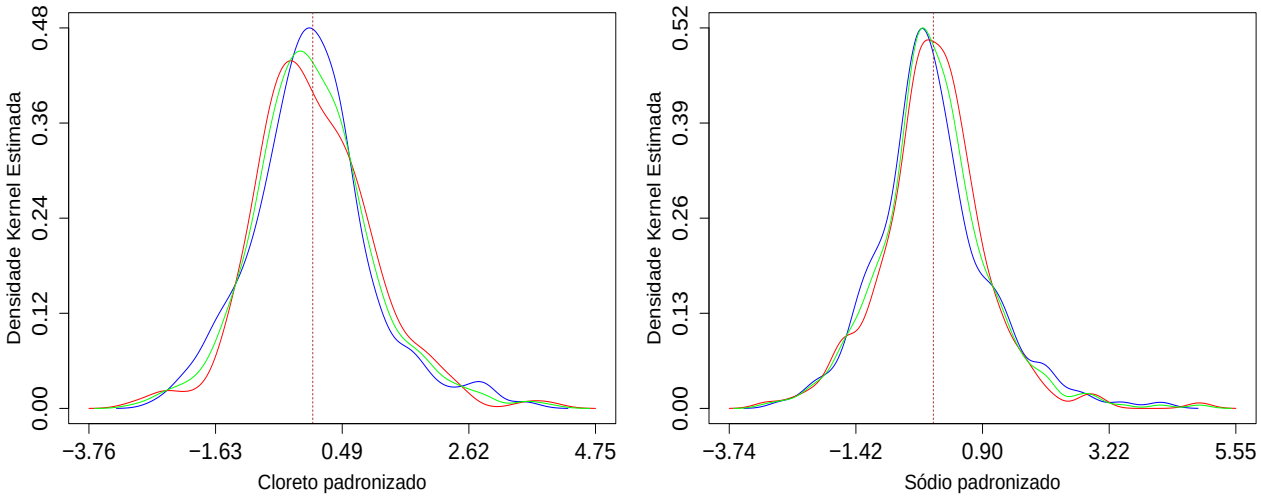

Figura 17 – Densidade Kernel estimada para as variáveis padronizadas Cloreto e Sódio (gram-negativo, gram-positivo, global, média).

Tabela 13 – Descrição das variáveis Cloreto e Sódio.

	Cloreto		Sódio	
Medidas	Gram-negativo	Gram-positivo	Gram-negativo	Gram-positivo
Válidos	238	211	238	211
%Válidos	85.921	85.772	85.921	85.772
Perdidos	39	35	39	35
%Perdidos	14.079	14.228	14.079	14.228
%Zeros	0.000	0.000	0.000	0.000
%(Z-P)	14.079	14.228	14.079	14.228
Média	108.870	108.536	136.706	136.573
D.P.	7.650	7.683	6.871	7.256
Mínimo	86.000	89.000	115.000	117.000
P10%	100.700	99.000	129.000	128.000
P25%	104.000	104.000	133.000	133.000
P50%	108.000	108.000	137.000	136.000
P75%	113.000	112.000	140.000	140.000
P90%	118.000	117.000	145.000	146.000
Máximo	139.000	136.000	171.000	166.000
ASC-ROC	0.48998		0.52241	
Distância L^2	0.00161		0.00183	

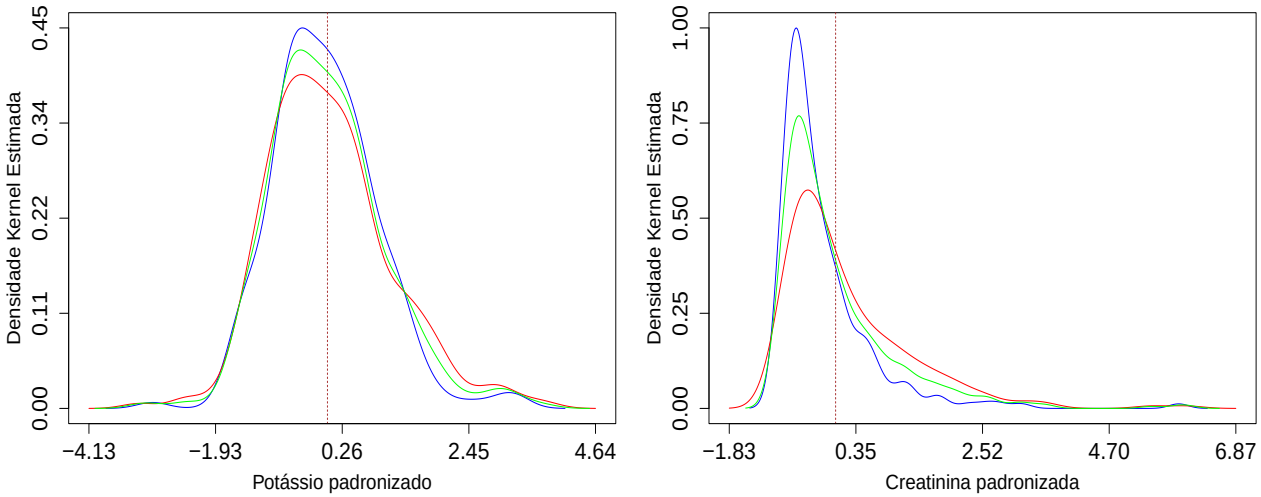

Figura 18 – Densidade Kernel estimada para as variáveis padronizadas Potássio e Creatinina (gram-negativo, gram-positivo, global, média).

Tabela 14 – Descrição das variáveis Potássio e Creatinina.

	Potássio		Creatinina	
Medidas	Gram-negativo	Gram-positivo	Gram-negativo	Gram-positivo
Válidos	238	210	228	203
%Válidos	85.921	85.366	82.310	82.520
Perdidos	39	36	49	43
%Perdidos	14.079	14.634	17.690	17.480
%Zeros	0.000	0.000	0.000	0.000
%(Z-P)	14.079	14.634	17.690	17.480
Média	4.008	3.951	1.818	1.206
D.P.	0.918	0.776	1.591	1.211
Mínimo	1.200	1.400	0.100	0.100
P10%	3.000	3.090	0.404	0.272
P25%	3.400	3.400	0.715	0.495
P50%	3.900	3.900	1.255	0.800
P75%	4.500	4.400	2.520	1.525
P90%	5.230	4.900	3.874	2.440
Máximo	7.200	6.800	10.300	10.100
ASC-ROC	0.49250		0.63635	
Distância L^2	0.00131		0.00520	

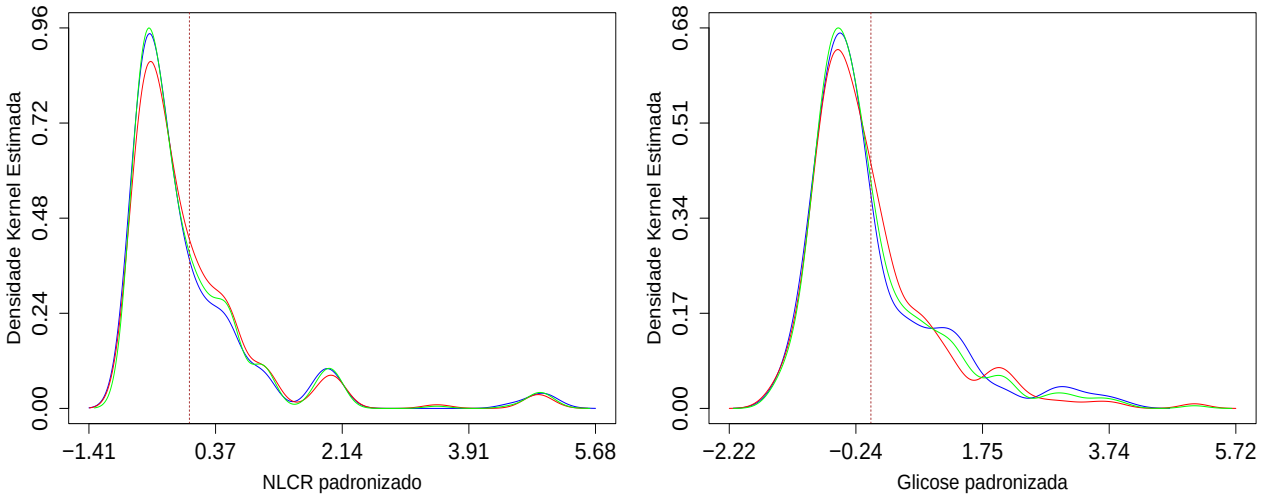

Figura 19 – Densidade Kernel estimada para as variáveis padronizadas NLCR e Glicose (gram-negativo, gram-positivo, global, média).

Tabela 15 – Descrição das variáveis NLCR e Glicose.

	NLCR		Glicose	
Medidas	Gram-negativo	Gram-positivo	Gram-negativo	Gram-positivo
Válidos	222	200	221	200
%Válidos	80.144	81.301	79.783	81.301
Perdidos	55	46	56	46
%Perdidos	19.856	18.699	20.217	18.699
%Zeros	0.000	0.000	0.000	0.000
%(Z-P)	19.856	18.699	20.217	18.699
Média	13.834	13.667	141.538	143.680
D.P.	16.336	17.530	73.067	76.900
Mínimo	0.056	0.011	25.000	36.000
P10%	1.622	1.622	78.000	77.900
P25%	3.833	3.381	94.000	95.750
P50%	8.450	7.682	123.000	120.500
P75%	18.350	17.700	165.000	174.500
P90%	30.667	31.032	235.000	244.700
Máximo	97.999	99.737	522.000	438.000
ASC-ROC	0.52227		0.50362	
Distância L^2	0.00090		0.00115	

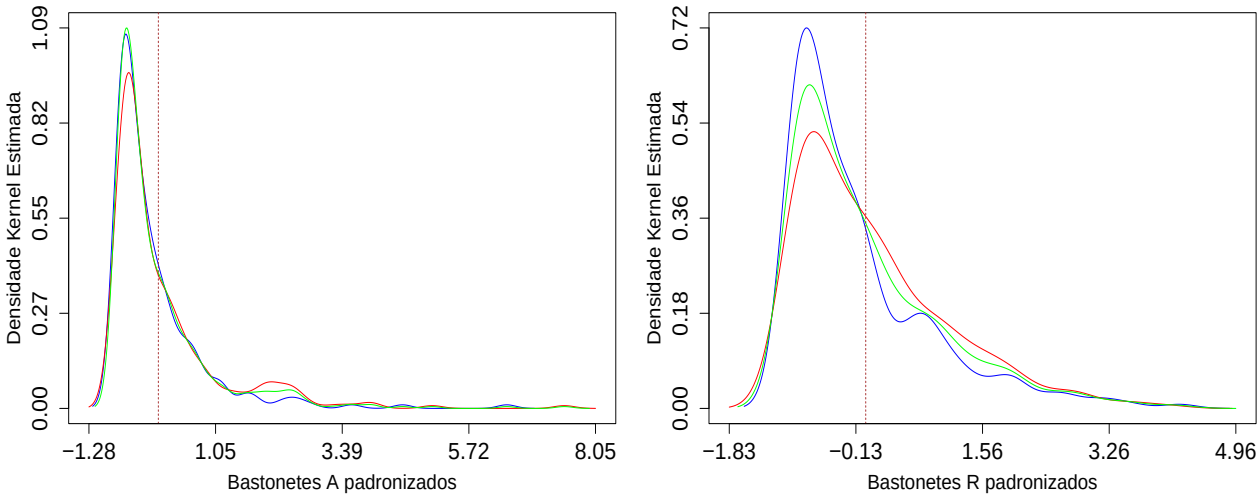

Figura 20 – Densidade Kernel estimada para as variáveis padronizadas Bastonetes A e Bastonetes R (gram-negativo, gram-positivo, global, média).

Tabela 16 – Descrição das variáveis Bastonetes A e Bastonetes R.

	Bastonetes A		Bastonetes R	
Medidas	Gram-negativo	Gram-positivo	Gram-negativo	Gram-positivo
Válidos	261	229	261	229
%Válidos	94.224	93.089	94.224	93.089
Perdidos	16	17	16	17
%Perdidos	5.776	6.911	5.776	6.911
%Zeros	12.996	16.260	12.996	16.260
%(Z-P)	18.773	23.171	18.773	23.171
Média	1897.487	1505.384	12.751	10.528
D.P.	2634.234	2182.176	12.331	11.964
Mínimo	0.000	0.000	0.000	0.000
P10%	0.000	0.000	0.000	0.000
P25%	324.000	153.000	3.000	2.000
P50%	856.000	821.000	10.000	6.000
P75%	2444.000	1974.000	19.000	14.000
P90%	5472.000	3579.800	30.000	26.000
Máximo	19953.000	17369.000	61.000	63.000
ASC-ROC	0.54234		0.56443	
Distância L^2	0.00169		0.00207	

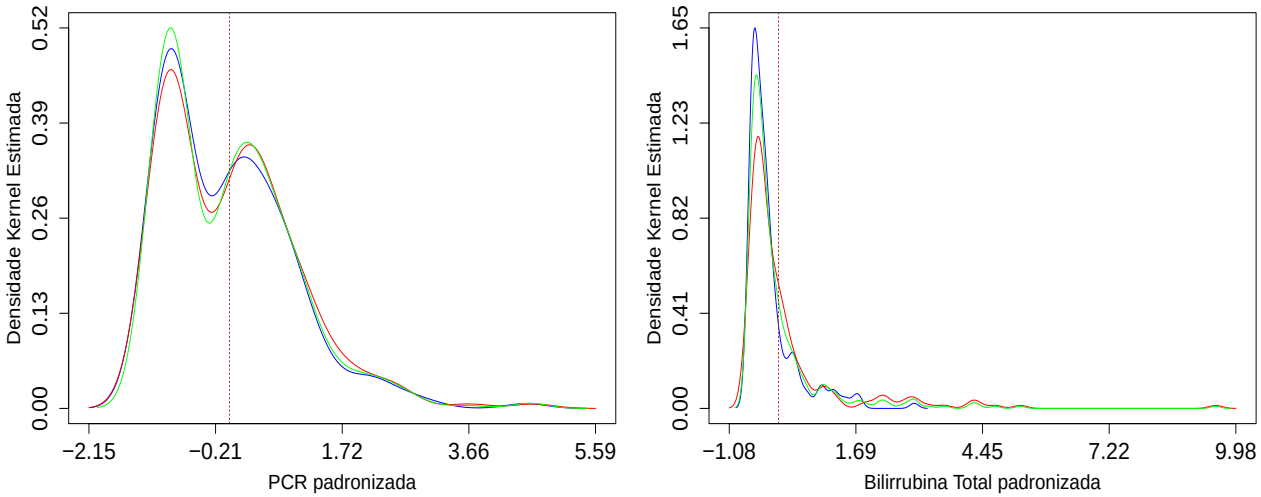

Figura 21 – Densidade Kernel estimada para as variáveis padronizadas PCR e Bilirrubina Total (gram-negativo, gram-positivo, global, média).

Tabela 17 – Descrição das variáveis PCR e Bilirrubina Total.

	PCR		Bilirrubina Total	
Medidas	Gram-negativo	Gram-positivo	Gram-negativo	Gram-positivo
Válidos	211	194	212	189
%Válidos	76.173	78.862	76.534	76.829
Perdidos	66	52	65	57
%Perdidos	23.827	21.138	23.466	23.171
%Zeros	0.000	0.000	0.361	0.813
%(Z-P)	23.827	21.138	23.827	23.984
Média	18.757	17.859	2.753	1.561
D.P.	15.041	14.542	4.314	1.919
Mínimo	0.500	0.500	0.000	0.000
P10%	2.800	2.520	0.200	0.100
P25%	5.700	5.425	0.600	0.400
P50%	18.100	16.650	1.150	0.900
P75%	26.900	25.750	2.800	1.800
P90%	36.700	34.300	6.480	3.720
Máximo	87.100	85.300	35.100	12.400
ASC-ROC	0.51605		0.58720	
Distância L^2	0.00052		0.00593	

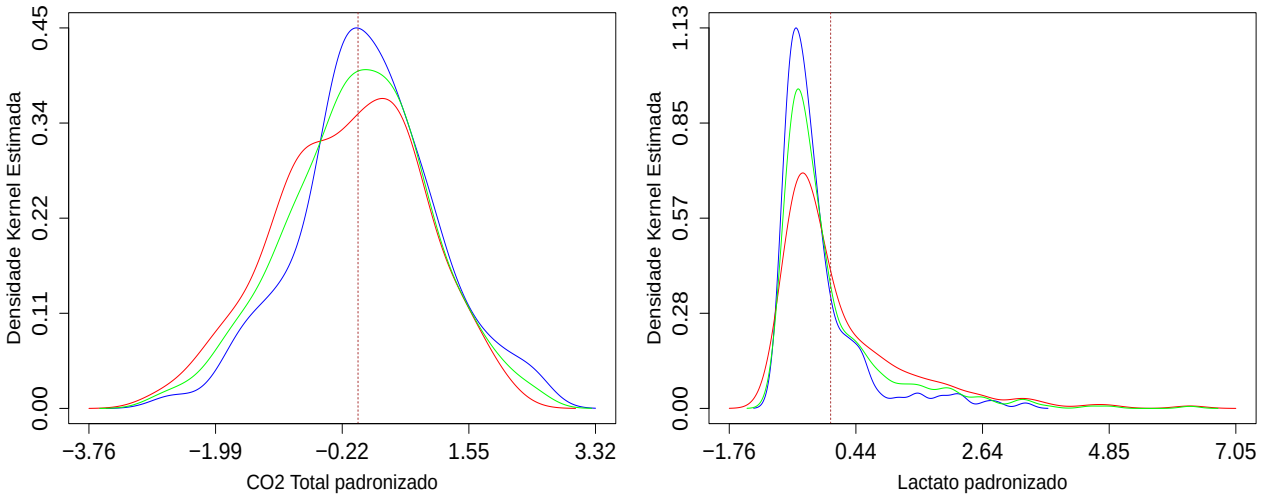

Figura 22 – Densidade Kernel estimada para as variáveis padronizadas CO2 Total e Lactato (gram-negativo, gram-positivo, global, média).

Tabela 18 – Descrição das variáveis CO2 Total e Lactato.

	CO2 Total		Lactato	
Medidas	Gram-negativo	Gram-positivo	Gram-negativo	Gram-positivo
Válidos	209	189	209	189
%Válidos	75.451	76.829	75.451	76.829
Perdidos	68	57	68	57
%Perdidos	24.549	23.171	24.549	23.171
%Zeros	0.000	0.000	0.000	0.000
%(Z-P)	24.549	23.171	24.549	23.171
Média	20.409	22.070	3.658	2.495
D.P.	6.456	6.201	3.309	2.114
Mínimo	3.200	4.100	0.340	0.400
P10%	11.940	14.200	1.100	0.900
P25%	16.000	18.400	1.500	1.200
P50%	21.000	22.000	2.400	1.800
P75%	25.200	26.000	4.800	2.800
P90%	28.360	29.980	7.960	4.600
Máximo	34.600	37.000	21.000	12.800
ASC-ROC	0.56531		0.61406	
Distância L^2	0.00168		0.00513	

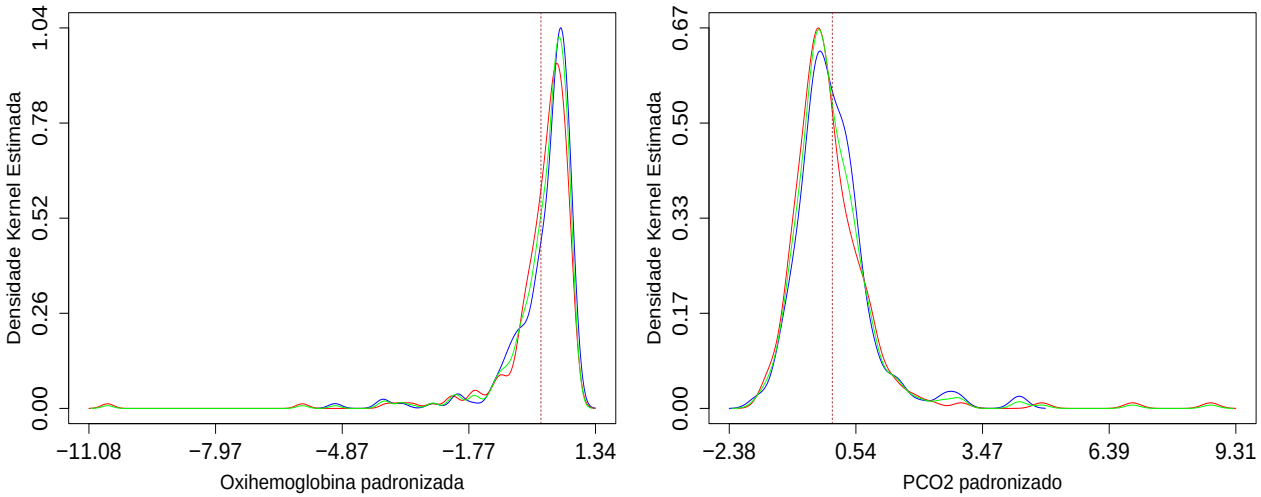

Figura 23 – Densidade Kernel estimada para as variáveis padronizadas Oxihemoglobina e pCO2 (gram-negativo, gram-positivo, global, média).

Tabela 19 – Descrição das variáveis Oxihemoglobina e pCO2.

	Oxihemoglobina		pCO2	
Medidas	Gram-negativo	Gram-positivo	Gram-negativo	Gram-positivo
Válidos	209	189	209	189
%Válidos	75.451	76.829	75.451	76.829
Perdidos	68	57	68	57
%Perdidos	24.549	23.171	24.549	23.171
%Zeros	0.000	0.000	0.000	0.000
%(Z-P)	24.549	23.171	24.549	23.171
Média	92.010	92.741	37.306	37.941
D.P.	7.630	5.863	16.954	13.423
Mínimo	19.600	57.800	13.300	10.200
P10%	86.420	87.100	24.000	24.600
P25%	91.100	91.400	29.200	30.200
P50%	94.100	94.700	34.300	36.200
P75%	95.600	96.100	42.200	43.100
P90%	96.700	97.100	50.700	51.220
Máximo	98.200	98.200	171.600	104.000
ASC-ROC	0.55159		0.53743	
Distância L^2	0.00396		0.00245	

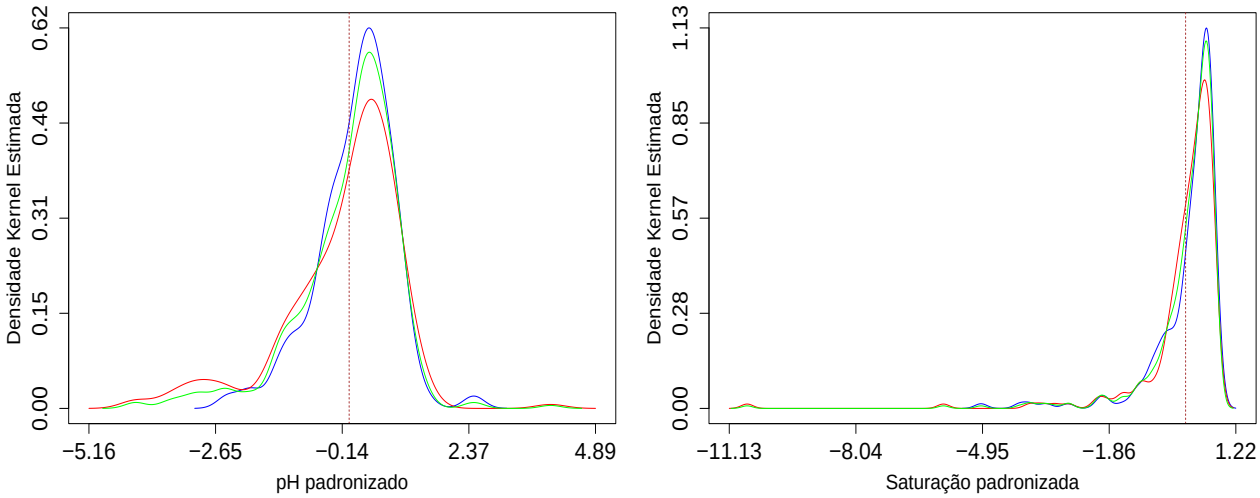

Figura 24 – Densidade Kernel estimada para as variáveis padronizadas pH e Saturação (gram-negativo, gram-positivo, global, média).

Tabela 20 – Descrição das variáveis pH e Saturação.

	pH		Saturação	
Medidas	Gram-negativo	Gram-positivo	Gram-negativo	Gram-positivo
Válidos	209	189	209	189
%Válidos	75.451	76.829	75.451	76.829
Perdidos	68	57	68	57
%Perdidos	24.549	23.171	24.549	23.171
%Zeros	0.000	0.000	0.000	0.000
%(Z-P)	24.549	23.171	24.549	23.171
Média	7.333	7.369	94.322	94.972
D.P.	0.171	0.114	7.809	5.932
Mínimo	6.720	6.990	20.000	59.800
P10%	7.140	7.218	89.100	89.380
P25%	7.250	7.310	93.600	93.900
P50%	7.380	7.386	96.200	97.000
P75%	7.440	7.440	98.000	98.400
P90%	7.490	7.490	99.020	99.100
Máximo	7.940	7.720	99.700	100.000
ASC-ROC	0.53873		0.53919	
Distância L^2	0.00237		0.00361	

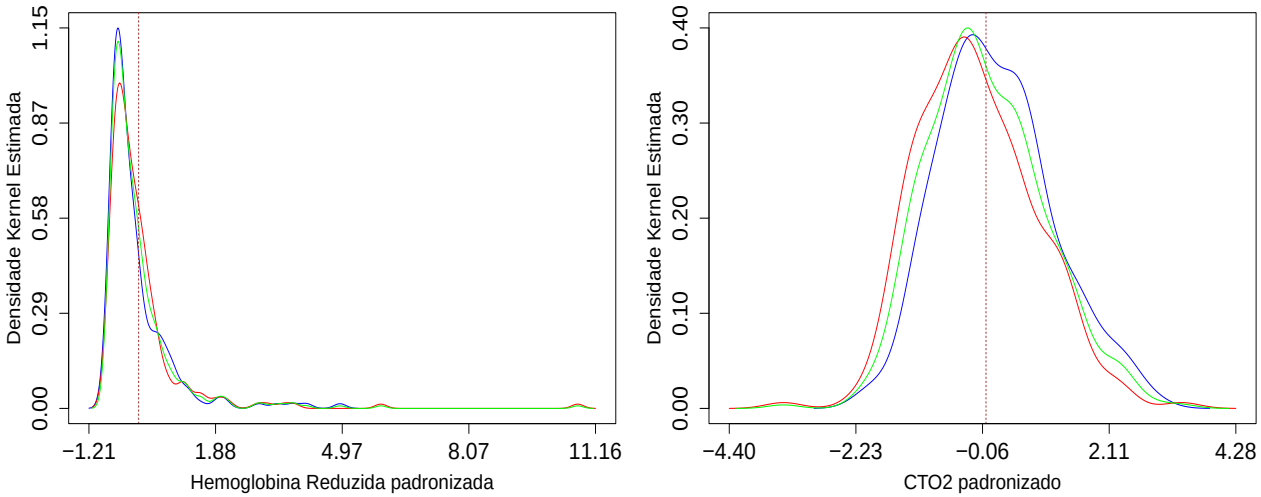

Figura 25 – Densidade Kernel estimada para as variáveis padronizadas Hemoglobina Reduzida e CtO2 (gram-negativo, gram-positivo, global, média).

Tabela 21 – Descrição das variáveis Hemoglobina Reduzida e CtO2.

	Hemoglobina Reduzida		CtO2	
Medidas	Gram-negativo	Gram-positivo	Gram-negativo	Gram-positivo
Válidos	209	189	208	189
%Válidos	75.451	76.829	75.090	76.829
Perdidos	68	57	69	57
%Perdidos	24.549	23.171	24.910	23.171
%Zeros	0.000	0.407	0.000	0.000
%(Z-P)	24.549	23.577	24.910	23.171
Média	5.544	4.911	13.502	14.664
D.P.	7.639	5.783	3.477	3.373
Mínimo	0.300	0.000	2.000	7.000
P10%	0.980	0.900	9.300	10.280
P25%	1.900	1.600	10.875	12.200
P50%	3.700	3.000	13.100	14.300
P75%	6.200	5.900	15.625	16.700
P90%	10.600	10.320	18.330	19.500
Máximo	78.400	38.800	25.700	24.200
ASC-ROC	0.53909		0.59800	
Distância L^2	0.00368		0.00231	

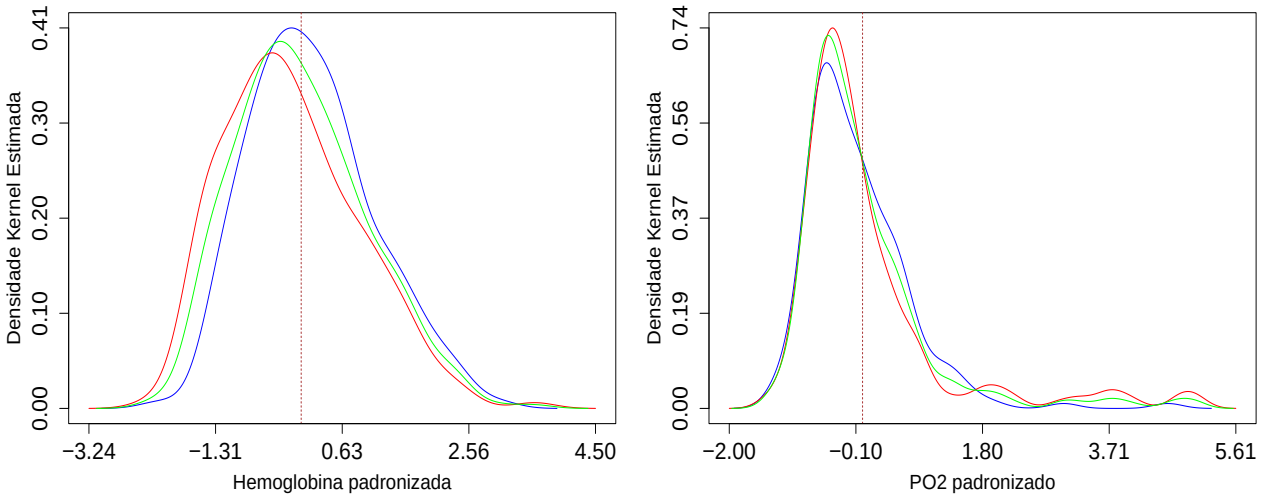

Figura 26 – Densidade Kernel estimada para as variáveis padronizadas Hemoglobina e pO2 (gram-negativo, gram-positivo, global, média).

Tabela 22 – Descrição das variáveis Hemoglobina e pO2.

	Hemoglobina		pO2	
Medidas	Gram-negativo	Gram-positivo	Gram-negativo	Gram-positivo
Válidos	209	188	208	189
%Válidos	75.451	76.423	75.090	76.829
Perdidos	68	58	69	57
%Perdidos	24.549	23.577	24.910	23.171
%Zeros	0.000	0.000	0.000	0.000
%(Z-P)	24.549	23.577	24.910	23.171
Média	10.290	11.073	105.592	96.924
D.P.	2.601	2.394	63.423	40.412
Mínimo	4.800	5.200	24.300	32.800
P10%	7.100	8.070	58.020	59.260
P25%	8.400	9.300	70.975	69.400
P50%	9.900	10.800	86.350	88.500
P75%	11.800	12.500	115.250	119.400
P90%	13.900	14.360	163.000	141.100
Máximo	19.700	18.300	372.700	348.300
ASC-ROC	0.59525		0.49668	
Distância L^2	0.00225		0.00157	

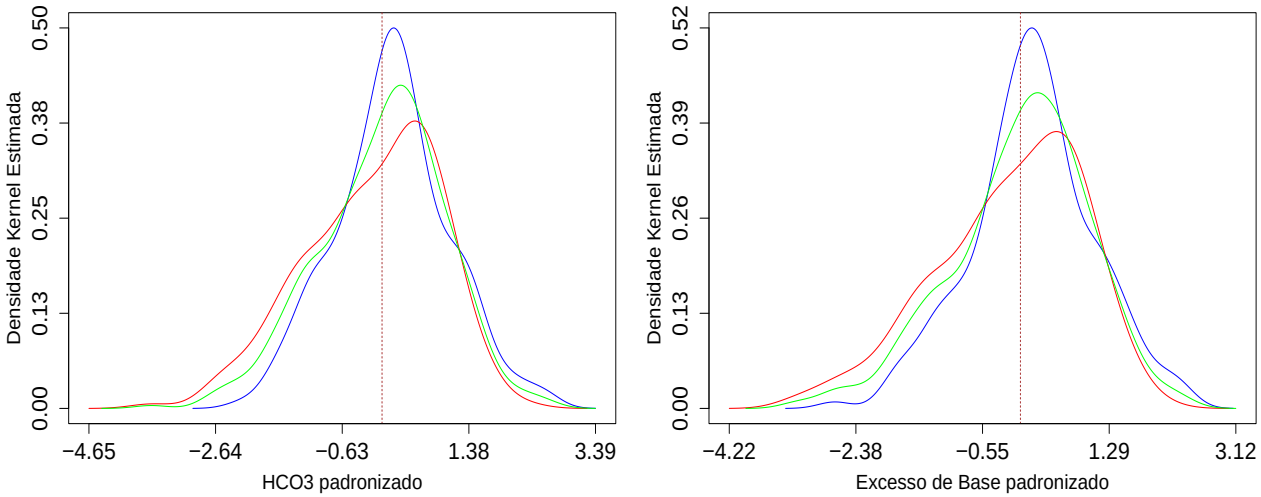

Figura 27 – Densidade Kernel estimada para as variáveis padronizadas HCO3 e Excesso de Base (gram-negativo, gram-positivo, global, média).

Tabela 23 – Descrição das variáveis HCO3 e Excesso de Base.

	HCO3		Excesso de Base	
Medidas	Gram-negativo	Gram-positivo	Gram-negativo	Gram-positivo
Válidos	208	189	209	189
%Válidos	75.090	76.829	75.451	76.829
Perdidos	69	57	68	57
%Perdidos	24.910	23.171	24.549	23.171
%Zeros	0.361	0.000	0.722	0.407
%(Z-P)	25.271	23.171	25.271	23.577
Média	19.929	21.668	-5.797	-3.547
D.P.	6.024	5.111	7.514	6.315
Mínimo	0.000	8.200	-27.500	-23.700
P10%	11.770	14.860	-16.340	-11.760
P25%	15.975	18.400	-10.500	-7.000
P50%	20.850	21.700	-4.600	-3.600
P75%	24.525	24.700	0.000	0.100
P90%	27.030	28.420	2.840	4.740
Máximo	33.600	35.500	9.600	12.000
ASC-ROC	0.56371		0.56740	
Distância L^2	0.00230		0.00223	

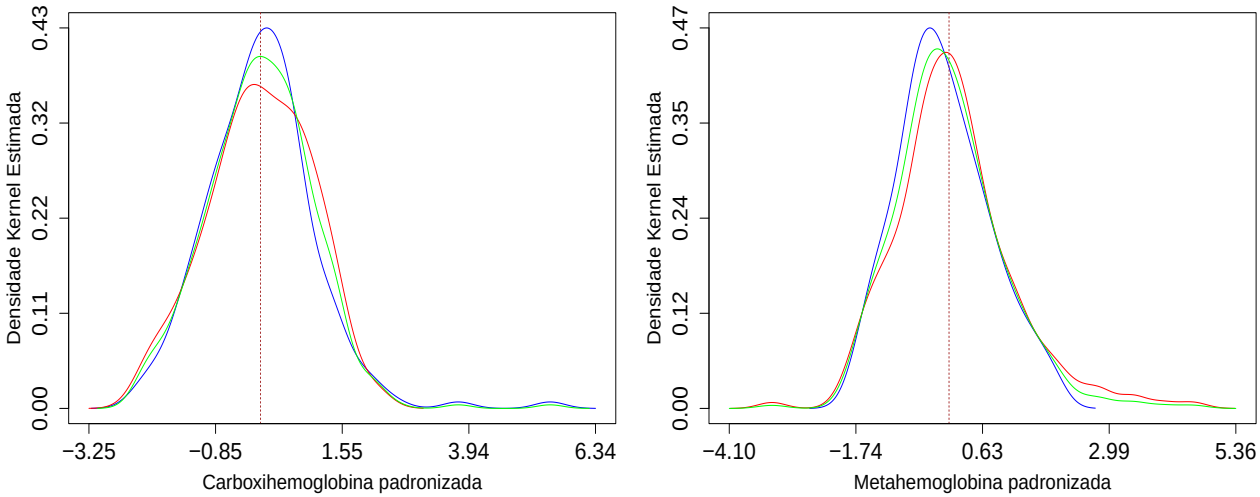

Figura 28 – Densidade Kernel estimada para as variáveis padronizadas Carboxihemoglobina e Metahemoglobina (gram-negativo, gram-positivo, global, média).

Tabela 24 – Descrição das variáveis Carboxihemoglobina e Metahemoglobina.

	Carboxihemoglobina		Metahemoglobina	
Medidas	Gram-negativo	Gram-positivo	Gram-negativo	Gram-positivo
Válidos	208	189	209	187
%Válidos	75.090	76.829	75.451	76.016
Perdidos	69	57	68	59
%Perdidos	24.910	23.171	24.549	23.984
%Zeros	0.361	0.407	0.361	0.000
%(Z-P)	25.271	23.577	24.910	23.984
Média	1.345	1.351	1.115	1.010
D.P.	0.560	0.593	0.595	0.451
Mínimo	0.000	0.000	-0.700	0.100
P10%	0.600	0.600	0.400	0.400
P25%	0.975	1.000	0.800	0.700
P50%	1.350	1.300	1.100	1.000
P75%	1.800	1.700	1.400	1.300
P90%	2.100	2.020	1.800	1.600
Máximo	2.600	4.500	3.500	2.100
ASC-ROC	0.51182		0.54791	
Distância L^2	0.00160		0.00167	

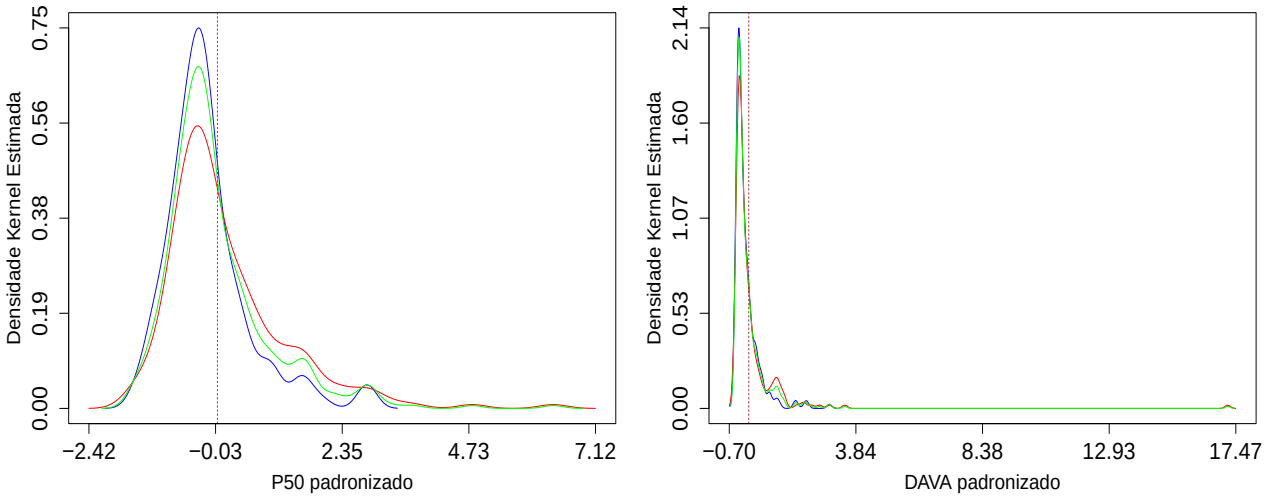

Figura 29 – Densidade Kernel estimada para as variáveis padronizadas p50 e DAV A (gram-negativo, gram-positivo, global, média).

Tabela 25 – Descrição das variáveis p50 e DAV A.

	p50		DAV A	
Medidas	Gram-negativo	Gram-positivo	Gram-negativo	Gram-positivo
Válidos	198	185	222	200
%Válidos	71.480	75.203	80.144	81.301
Perdidos	79	61	55	46
%Perdidos	28.520	24.797	19.856	18.699
%Zeros	0.000	0.000	11.552	14.228
%(Z-P)	28.520	24.797	31.408	32.927
Média	28.362	26.454	2578.158	1657.555
D.P.	6.194	4.350	6850.954	2465.027
Mínimo	18.600	18.890	0.000	0.000
P10%	22.808	22.094	0.000	0.000
P25%	24.523	23.800	297.000	152.500
P50%	26.510	25.660	917.000	822.000
P75%	30.755	28.050	2549.250	2126.750
P90%	36.566	31.870	6758.200	4006.400
Máximo	61.930	42.920	92581.000	17369.000
ASC-ROC	0.58972		0.54222	
Distância L^2	0.00309		0.00502	

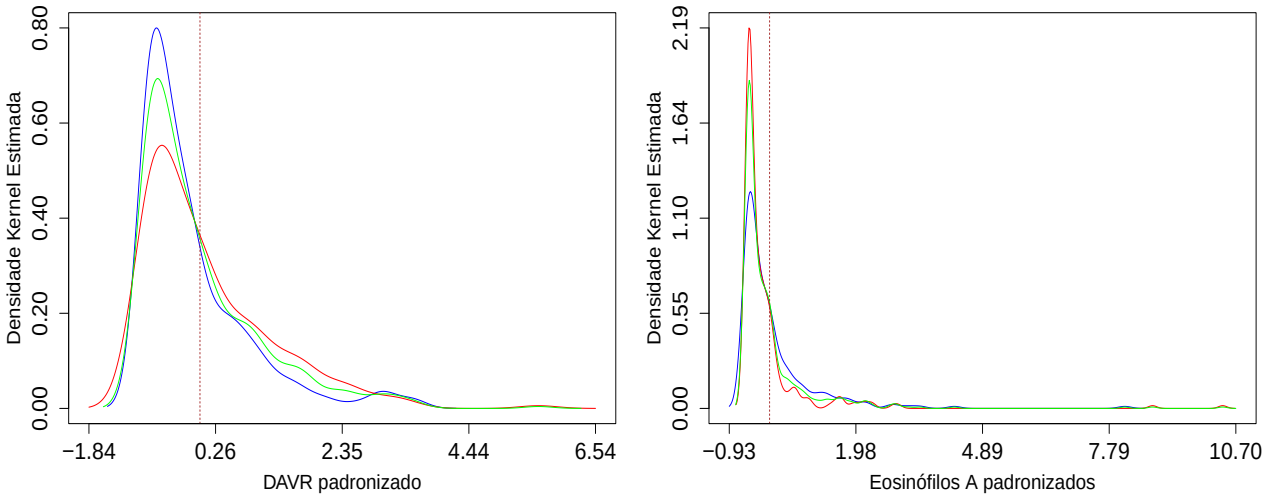

Figura 30 – Densidade Kernel estimada para as variáveis padronizadas DAV R e Eosinófilos A (gram-negativo, gram-positivo, global, média).

Tabela 26 – Descrição das variáveis DAV R e Eosinófilos A.

	DAV R		Eosinófilos A	
Medidas	Gram-negativo	Gram-positivo	Gram-negativo	Gram-positivo
Válidos	222	200	261	229
%Válidos	80.144	81.301	94.224	93.089
Perdidos	55	46	16	17
%Perdidos	19.856	18.699	5.776	6.911
%Zeros	11.552	14.228	43.321	35.366
%(Z-P)	31.408	32.927	49.097	42.276
Média	14.527	11.140	153.211	207.762
D.P.	15.383	13.253	389.340	352.007
Mínimo	0.000	0.000	0.000	0.000
P10%	0.000	0.000	0.000	0.000
P25%	3.000	2.000	0.000	0.000
P50%	10.000	7.000	35.100	93.300
P75%	23.000	15.250	160.500	245.200
P90%	37.000	27.100	387.200	595.200
Máximo	94.000	65.000	4056.000	3217.500
ASC-ROC	0.56779		0.57417	
Distância L^2	0.00296		0.00844	

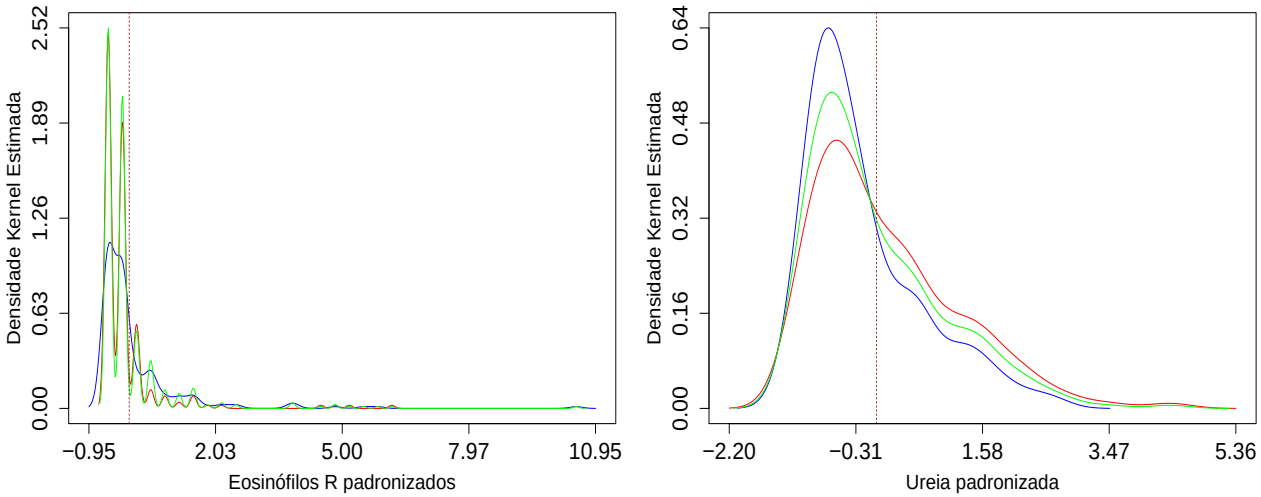

Figura 31 – Densidade Kernel estimada para as variáveis padronizadas Eosinófilos R e Ureia (gram-negativo, gram-positivo, global, média).

Tabela 27 – Descrição das variáveis Eosinófilos R e Ureia.

	Eosinófilos R		Ureia	
Medidas	Gram-negativo	Gram-positivo	Gram-negativo	Gram-positivo
Válidos	261	229	143	113
%Válidos	94.224	93.089	51.625	45.935
Perdidos	16	17	134	133
%Perdidos	5.776	6.911	48.375	54.065
%Zeros	43.321	35.366	0.000	0.000
%(Z-P)	49.097	42.276	48.375	54.065
Média	1.138	1.869	92.049	71.841
D.P.	2.367	3.560	64.933	51.872
Mínimo	0.000	0.000	10.000	8.000
P10%	0.000	0.000	26.000	23.400
P25%	0.000	0.000	42.000	34.000
P50%	1.000	1.000	73.000	57.000
P75%	1.000	2.000	121.500	98.000
P90%	2.000	5.000	180.800	154.600
Máximo	20.000	33.000	347.000	243.000
ASC-ROC	0.56726		0.59017	
Distância L^2	0.01341		0.00255	

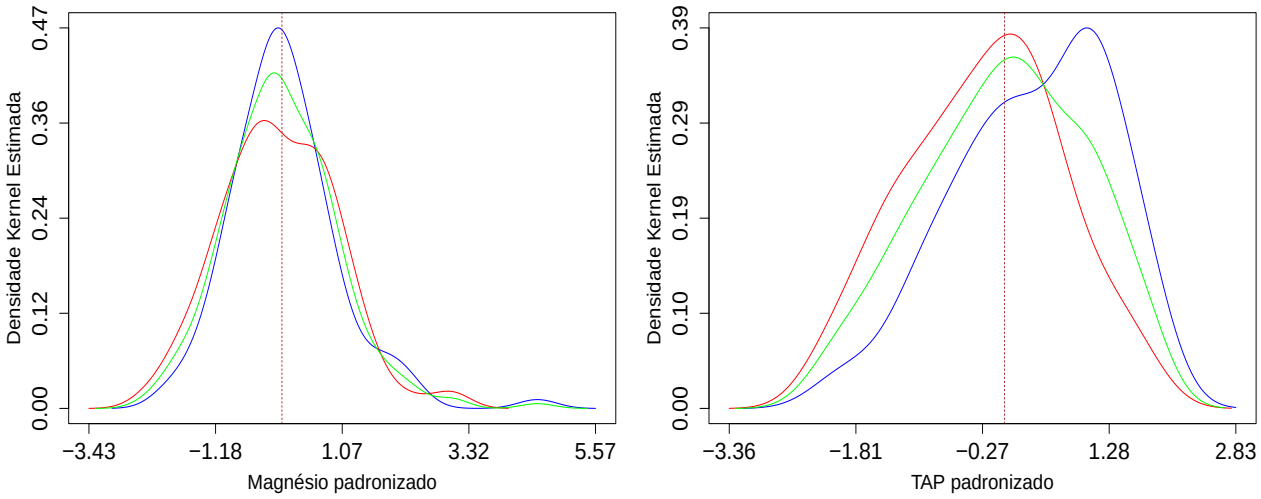

Figura 32 – Densidade Kernel estimada para as variáveis padronizadas Magnésio e TAP (gram-negativo, gram-positivo, global, média).

Tabela 28 – Descrição das variáveis Magnésio e TAP.

	Magnésio		TAP	
Medidas	Gram-negativo	Gram-positivo	Gram-negativo	Gram-positivo
Válidos	121	105	98	73
%Válidos	43.682	42.683	35.379	29.675
Perdidos	156	141	179	173
%Perdidos	56.318	57.317	64.621	70.325
%Zeros	0.000	0.000	0.000	0.000
%(Z-P)	56.318	57.317	64.621	70.325
Média	1.894	1.935	55.935	69.085
D.P.	0.473	0.446	21.343	21.175
Mínimo	0.800	1.000	10.000	14.800
P10%	1.300	1.500	27.250	39.760
P25%	1.600	1.600	39.425	54.000
P50%	1.900	1.900	58.050	70.000
P75%	2.200	2.200	70.750	86.000
P90%	2.500	2.400	82.210	95.180
Máximo	3.300	4.000	100.000	100.000
ASC-ROC	0.51669		0.67179	
Distância L^2	0.00218		0.00298	

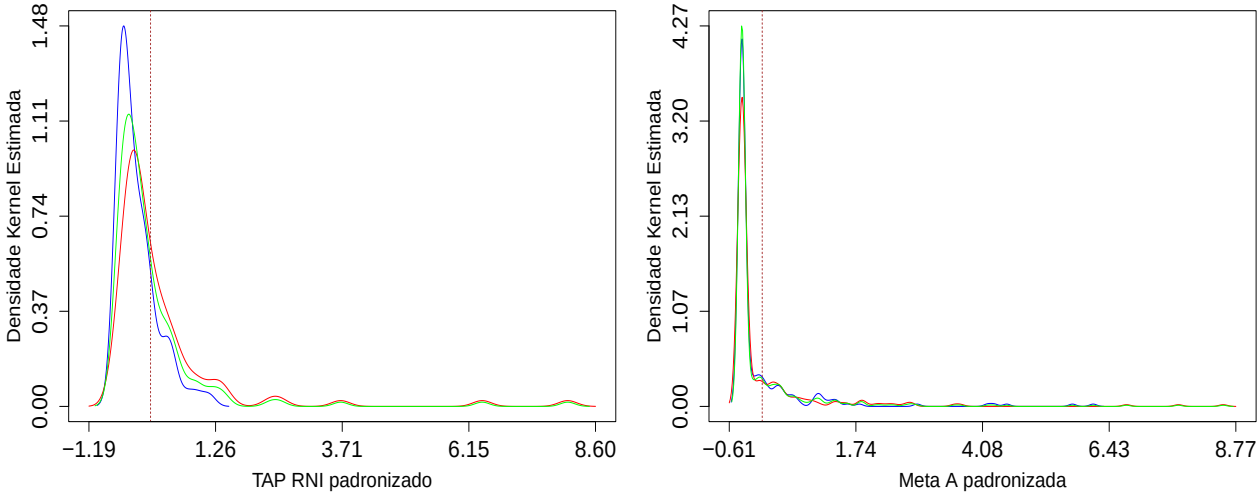

Figura 33 – Densidade Kernel estimada para as variáveis padronizadas TAP RNI e Meta A (gram-negativo, gram-positivo, global, média).

Tabela 29 – Descrição das variáveis TAP RNI e Meta A.

	TAP RNI		Meta A	
Medidas	Gram-negativo	Gram-positivo	Gram-negativo	Gram-positivo
Válidos	98	73	261	229
%Válidos	35.379	29.675	94.224	93.089
Perdidos	179	173	16	17
%Perdidos	64.621	70.325	5.776	6.911
%Zeros	0.000	0.000	63.177	63.415
%(Z-P)	64.621	70.325	68.953	70.325
Média	1.689	1.315	142.751	118.044
D.P.	1.010	0.313	377.573	312.177
Mínimo	1.000	1.000	0.000	0.000
P10%	1.110	1.032	0.000	0.000
P25%	1.223	1.090	0.000	0.000
P50%	1.409	1.210	0.000	0.000
P75%	1.765	1.450	123.000	102.000
P90%	2.368	1.776	402.000	330.400
Máximo	8.070	2.450	3106.000	2267.000
ASC-ROC	0.67445		0.49073	
Distância L^2	0.00733		0.00376	

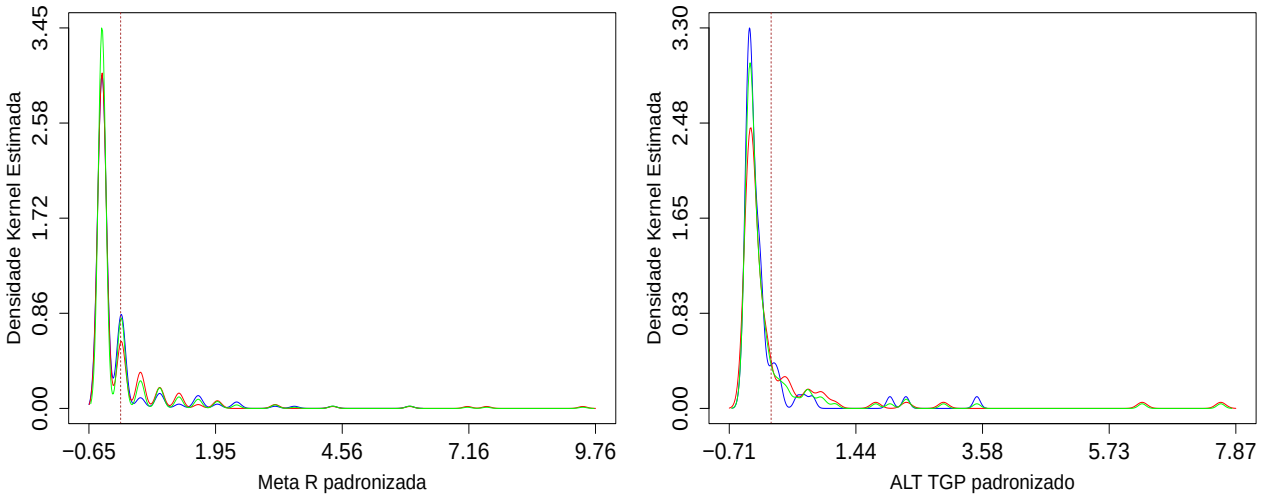

Figura 34 – Densidade Kernel estimada para as variáveis padronizadas Meta R e ALT TGP (gram-negativo, gram-positivo, global, média).

Tabela 30 – Descrição das variáveis Meta R e ALT TGP.

	Meta R		ALT TGP	
Medidas	Gram-negativo	Gram-positivo	Gram-negativo	Gram-positivo
Válidos	261	229	88	68
%Válidos	94.224	93.089	31.769	27.642
Perdidos	16	17	189	178
%Perdidos	5.776	6.911	68.231	72.358
%Zeros	63.177	63.415	0.000	0.000
%(Z-P)	68.953	70.325	68.231	72.358
Média	1.080	0.843	124.852	81.676
D.P.	2.893	2.037	248.986	133.226
Mínimo	0.000	0.000	12.000	6.000
P10%	0.000	0.000	22.700	22.000
P25%	0.000	0.000	31.750	28.000
P50%	0.000	0.000	46.500	41.000
P75%	1.000	1.000	96.000	69.000
P90%	3.000	3.000	248.800	134.100
Máximo	25.000	16.000	1684.000	828.000
ASC-ROC	0.48741		0.56375	
Distância L^2	0.00307		0.00519	

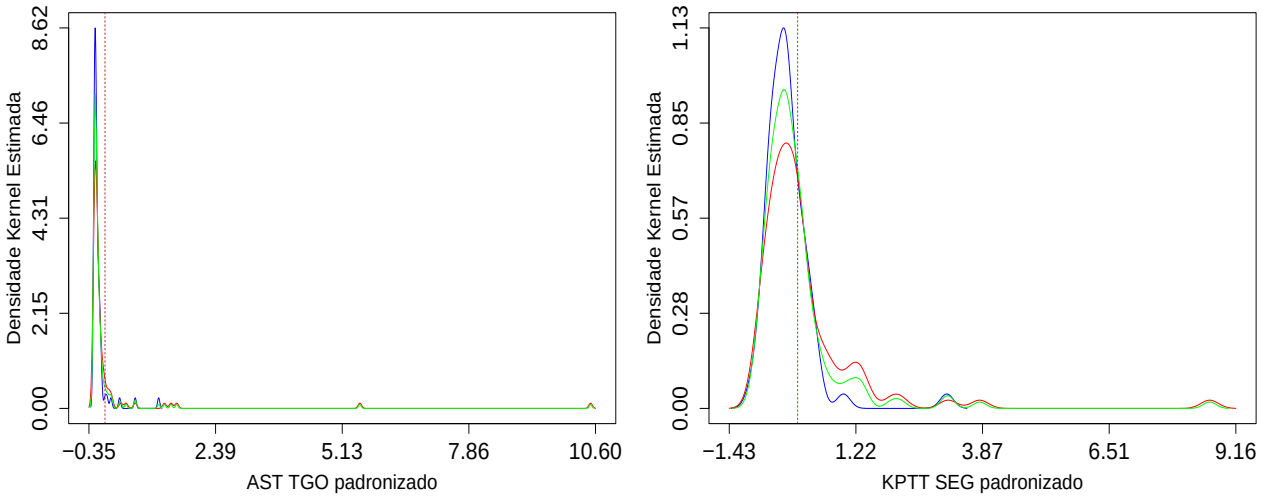

Figura 35 – Densidade Kernel estimada para as variáveis padronizadas AST TGO e KPTT SEG (gram-negativo, gram-positivo, global, média).

Tabela 31 – Descrição das variáveis AST TGO e KPTT SEG.

	AST TGO		KPTT SEG	
Medidas	Gram-negativo	Gram-positivo	Gram-negativo	Gram-positivo
Válidos	87	68	89	66
%Válidos	31.408	27.642	32.130	26.829
Perdidos	190	178	188	180
%Perdidos	68.592	72.358	67.870	73.171
%Zeros	0.000	0.000	0.000	0.000
%(Z-P)	68.592	72.358	67.870	73.171
Média	271.379	87.515	44.216	35.753
D.P.	980.533	155.403	29.239	12.951
Mínimo	11.000	15.000	19.400	17.400
P10%	21.600	18.700	24.500	25.150
P25%	29.000	24.750	30.600	28.575
P50%	58.000	43.500	37.000	33.900
P75%	125.000	87.750	46.600	40.200
P90%	293.200	127.300	69.120	46.100
Máximo	8013.000	1055.000	247.500	115.500
ASC-ROC	0.58426		0.59244	
Distância L^2	0.01344		0.00499	

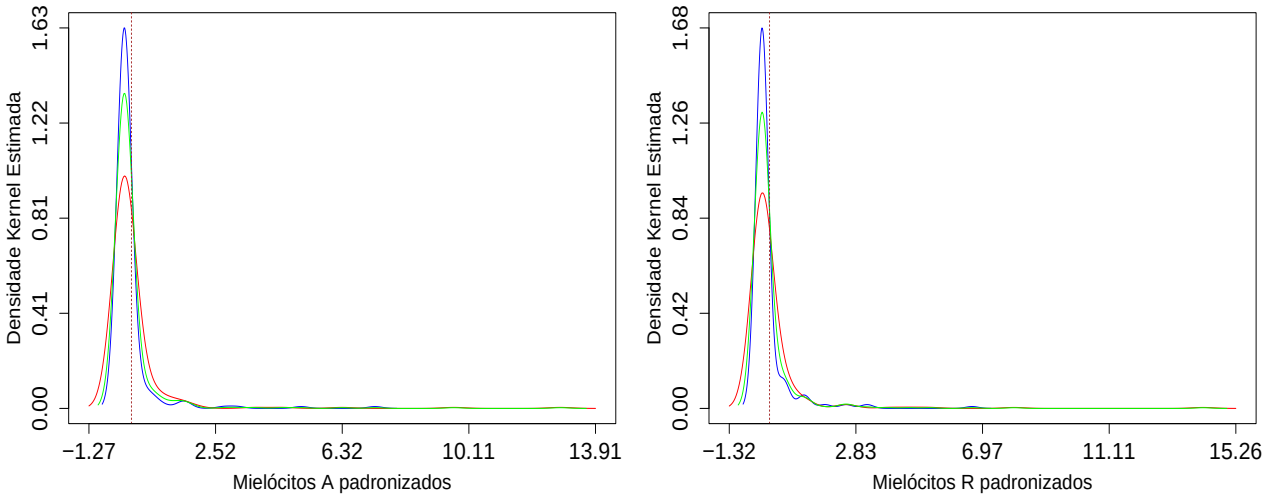

Figura 36 – Densidade Kernel estimada para as variáveis padronizadas Mielócitos A e Mielócitos R (gram-negativo, gram-positivo, global, média).

Tabela 32 – Descrição das variáveis Mielócitos A e Mielócitos R.

	Mielócitos A		Mielócitos R	
Medidas	Gram-negativo	Gram-positivo	Gram-negativo	Gram-positivo
Válidos	261	229	261	229
%Válidos	94.224	93.089	94.224	93.089
Perdidos	16	17	16	17
%Perdidos	5.776	6.911	5.776	6.911
%Zeros	79.061	81.301	79.061	81.301
%(Z-P)	84.838	88.211	84.838	88.211
Média	67.061	40.978	0.425	0.275
D.P.	299.220	182.089	1.763	0.995
Mínimo	0.000	0.000	0.000	0.000
P10%	0.000	0.000	0.000	0.000
P25%	0.000	0.000	0.000	0.000
P50%	0.000	0.000	0.000	0.000
P75%	0.000	0.000	0.000	0.000
P90%	165.000	70.400	1.000	1.000
Máximo	3286.000	1889.000	21.000	10.000
ASC-ROC	0.48173		0.48339	
Distância L^2	0.01078		0.01327	

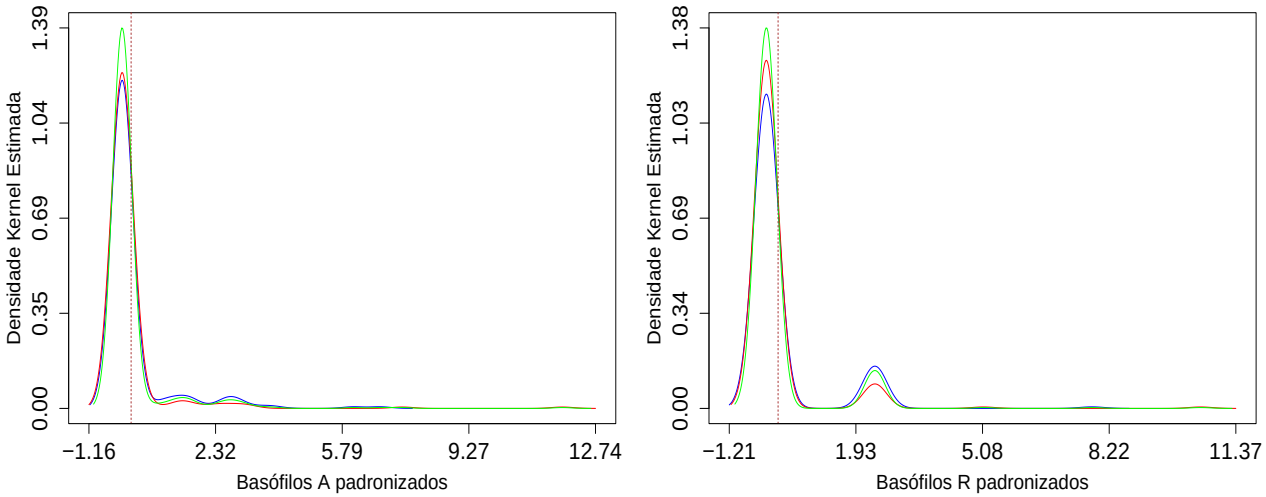

Figura 37 – Densidade Kernel estimada para as variáveis padronizadas Basófilos A e Basófilos R (gram-negativo, gram-positivo, global, média).

Tabela 33 – Descrição das variáveis Basófilos A e Basófilos R.

	Basófilos A		Basófilos R	
Medidas	Gram-negativo	Gram-positivo	Gram-negativo	Gram-positivo
Válidos	261	229	261	229
%Válidos	94.224	93.089	94.224	93.089
Perdidos	16	17	16	17
%Perdidos	5.776	6.911	5.776	6.911
%Zeros	87.365	81.707	87.365	81.707
%(Z-P)	93.141	88.618	93.141	88.618
Média	10.285	15.473	0.088	0.131
D.P.	52.720	49.856	0.367	0.375
Mínimo	0.000	0.000	0.000	0.000
P10%	0.000	0.000	0.000	0.000
P25%	0.000	0.000	0.000	0.000
P50%	0.000	0.000	0.000	0.000
P75%	0.000	0.000	0.000	0.000
P90%	0.000	60.920	0.000	1.000
Máximo	621.100	364.200	4.000	3.000
ASC-ROC	0.52500		0.52443	
Distância L^2	0.00138		0.00220	

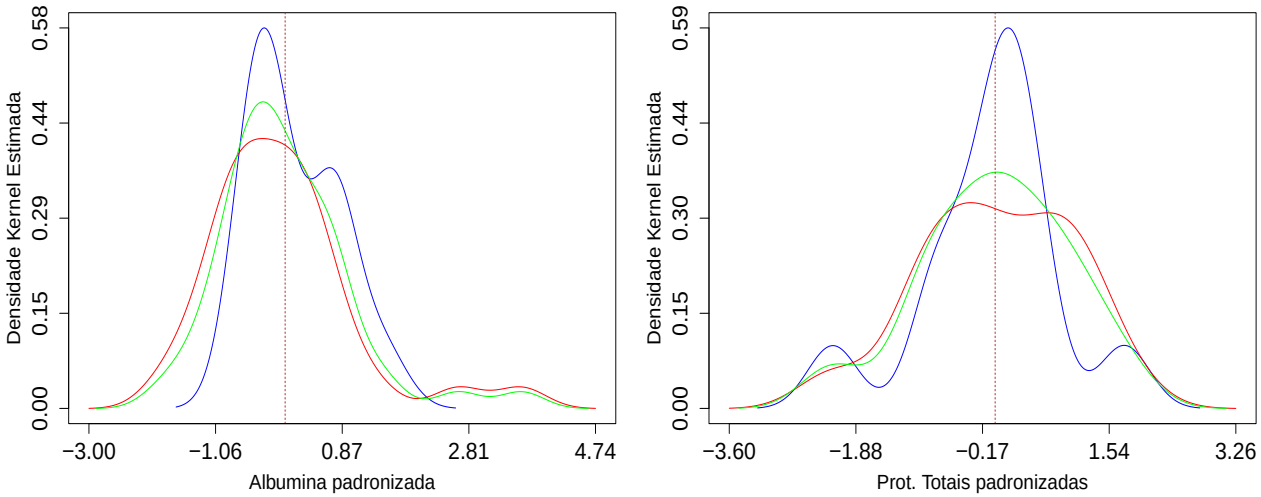

Figura 38 – Densidade Kernel estimada para as variáveis padronizadas Albumina e Proteínas Totais (gram-negativo, gram-positivo, global, média).

Tabela 34 – Descrição das variáveis Albumina e Proteínas Totais.

	Albumina		Proteínas Totais	
Medidas	Gram-negativo	Gram-positivo	Gram-negativo	Gram-positivo
Válidos	33	13	31	12
%Válidos	11.913	5.285	11.191	4.878
Perdidos	244	233	246	234
%Perdidos	88.087	94.715	88.809	95.122
%Zeros	0.000	0.000	0.000	0.000
%(Z-P)	88.087	94.715	88.809	95.122
Média	2.221	2.354	4.835	4.758
D.P.	0.686	0.418	0.993	0.917
Mínimo	1.100	1.900	2.700	2.700
P10%	1.520	1.920	3.800	4.010
P25%	1.800	2.000	4.100	4.400
P50%	2.200	2.200	4.800	4.900
P75%	2.500	2.700	5.600	5.200
P90%	2.780	2.780	6.100	5.380
Máximo	4.500	3.200	6.600	6.500
ASC-ROC	0.61422		0.48387	
Distância L^2	0.00348		0.00361	

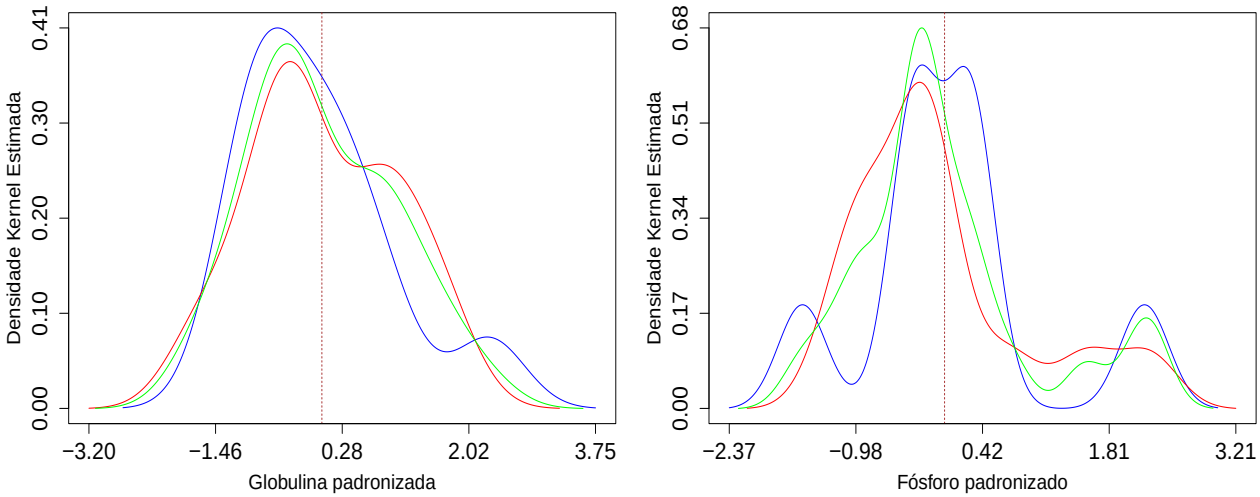

Figura 39 – Densidade Kernel estimada para as variáveis padronizadas Globulina e Fósforo (gram-negativo, gram-positivo, global, média).

Tabela 35 – Descrição das variáveis Globulina e Fósforo.

	Globulina		Fósforo	
Medidas	Gram-negativo	Gram-positivo	Gram-negativo	Gram-positivo
Válidos	30	11	13	8
%Válidos	10.830	4.472	4.693	3.252
Perdidos	247	235	264	238
%Perdidos	89.170	95.528	95.307	96.748
%Zeros	0.000	0.000	0.000	0.000
%(Z-P)	89.170	95.528	95.307	96.748
Média	2.643	2.582	4.577	4.787
D.P.	0.731	0.752	2.027	2.134
Mínimo	1.300	1.700	2.200	1.500
P10%	1.780	1.900	2.740	3.250
P25%	2.200	2.050	3.100	4.000
P50%	2.500	2.500	4.100	4.600
P75%	3.200	2.900	4.700	5.225
P90%	3.700	3.300	7.460	6.440
Máximo	4.000	4.300	9.200	9.100
ASC-ROC	0.45303		0.54808	
Distância L^2	0.00196		0.00375	

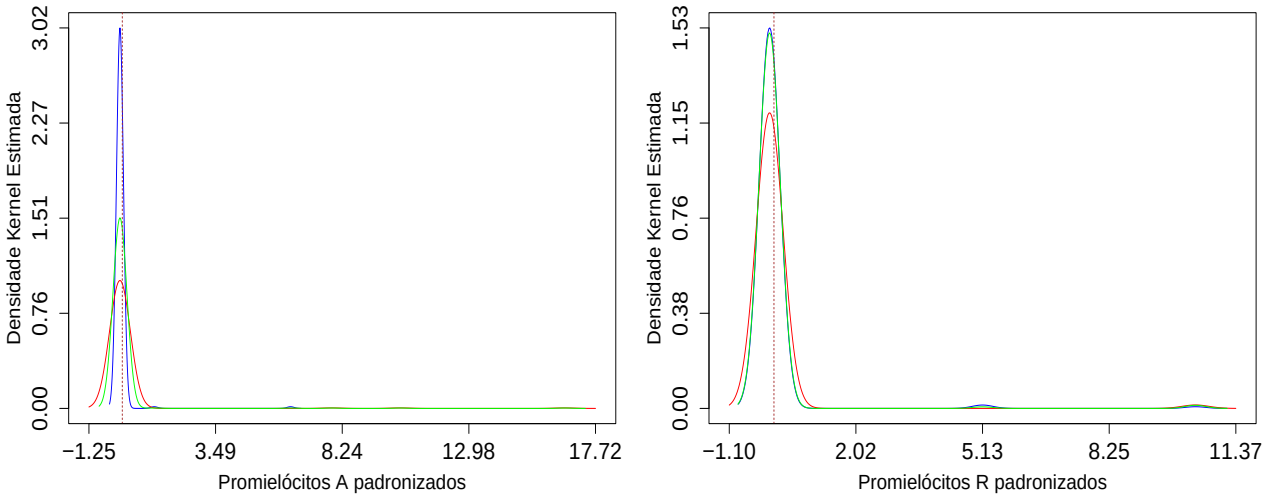

Figura 40 – Densidade Kernel estimada para as variáveis padronizadas Promielócitos A e Promielócitos R (gram-negativo, gram-positivo, global, média).

Tabela 36 – Descrição das variáveis Promielócitos A e Promielócitos R.

	Promielócitos A		Promielócitos R	
Medidas	Gram-negativo	Gram-positivo	Gram-negativo	Gram-positivo
Válidos	261	229	261	229
%Válidos	94.224	93.089	94.224	93.089
Perdidos	16	17	16	17
%Perdidos	5.776	6.911	5.776	6.911
%Zeros	93.141	91.870	93.141	91.870
%(Z-P)	98.917	98.780	98.917	98.780
Média	15.904	4.013	0.023	0.017
D.P.	154.909	50.916	0.214	0.161
Mínimo	0.000	0.000	0.000	0.000
P10%	0.000	0.000	0.000	0.000
P25%	0.000	0.000	0.000	0.000
P50%	0.000	0.000	0.000	0.000
P75%	0.000	0.000	0.000	0.000
P90%	0.000	0.000	0.000	0.000
Máximo	1970.000	756.000	2.000	2.000
ASC-ROC	0.50073		0.50075	
Distância L^2	0.03415		0.00510	

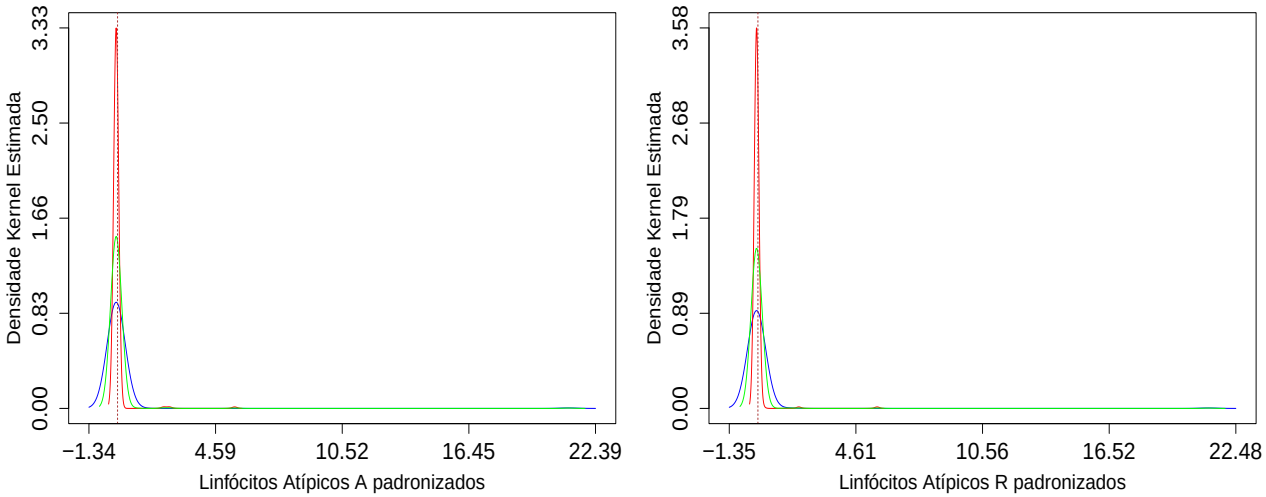

Figura 41 – Densidade Kernel estimada para as variáveis padronizadas Linfócitos Atípicos A e Linfócitos Atípicos R (gram-negativo, gram-positivo, global, média).

Tabela 37 – Descrição das variáveis Linfócitos Atípicos A e Linfócitos Atípicos R.

	Linfócitos Atípicos A		Linfócitos Atípicos R	
Medidas	Gram-negativo	Gram-positivo	Gram-negativo	Gram-positivo
Válidos	261	229	261	228
%Válidos	94.224	93.089	94.224	92.683
Perdidos	16	17	16	18
%Perdidos	5.776	6.911	5.776	7.317
%Zeros	93.141	92.276	93.141	91.870
%(Z-P)	98.917	99.187	98.917	99.187
Média	3.329	8.146	0.107	0.351
D.P.	33.831	118.529	1.311	4.977
Mínimo	0.000	0.000	0.000	0.000
P10%	0.000	0.000	0.000	0.000
P25%	0.000	0.000	0.000	0.000
P50%	0.000	0.000	0.000	0.000
P75%	0.000	0.000	0.000	0.000
P90%	0.000	0.000	0.000	0.000
Máximo	469.500	1792.500	20.000	75.000
ASC-ROC	0.49862		0.49866	
Distância L^2	0.04862		0.05191	

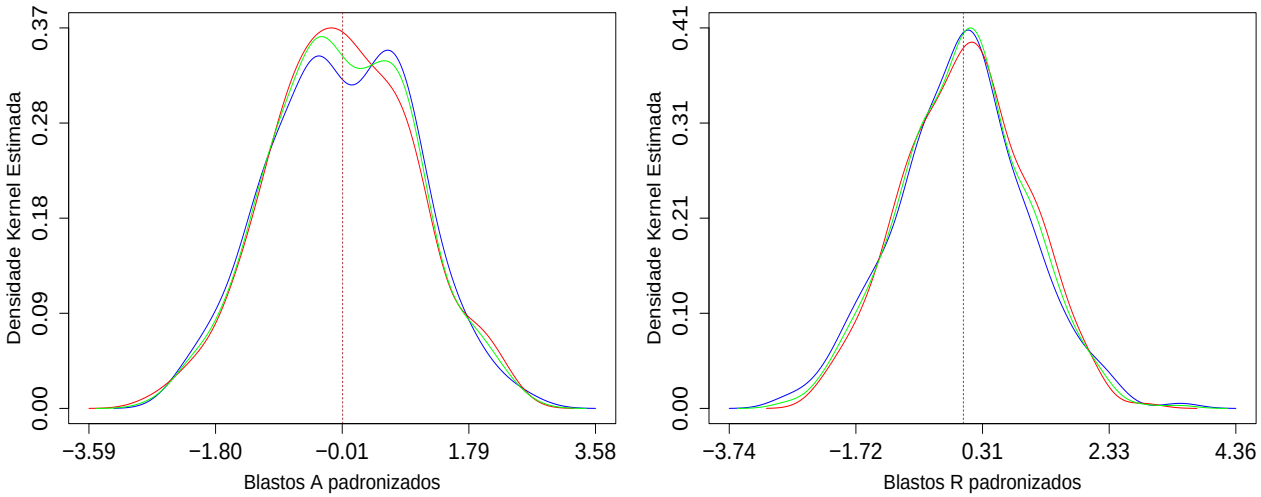

Figura 42 – Densidade Kernel estimada para as variáveis padronizadas Blastos A e Blastos R (gram-negativo, gram-positivo, global, média).

Tabela 38 – Descrição das variáveis Blastos A e Blastos R.

	Blastos A		Blastos R	
Medidas	Gram-negativo	Gram-positivo	Gram-negativo	Gram-positivo
Válidos	261	229	261	229
%Válidos	94.224	93.089	94.224	93.089
Perdidos	16	17	16	17
%Perdidos	5.776	6.911	5.776	6.911
%Zeros	93.141	93.089	93.141	93.089
%(Z-P)	98.917	100.000	98.917	100.000
Média	342.216	0.000	0.372	0.000
D.P.	5486.715	0.000	5.582	0.000
Mínimo	0.000	0.000	0.000	0.000
P10%	0.000	0.000	0.000	0.000
P25%	0.000	0.000	0.000	0.000
P50%	0.000	0.000	0.000	0.000
P75%	0.000	0.000	0.000	0.000
P90%	0.000	0.000	0.000	0.000
Máximo	88641.000	0.000	90.000	0.000
ASC-ROC	0.49425		0.49425	
Distância L^2	0.03872		0.03772	

APÊNDICE B

Descrição das Variáveis Laboratoriais

Tabela 39 – Descrição das variáveis laboratoriais utilizadas no estudo.

Variável	Descrição	Unidade
Albumina	Proteína plasmática sintetizada pelo fígado; mantém pressão oncótica	g/dL
ALT (TGP)	Enzima hepática (alanina aminotransferase), marcador de lesão hepática	U/L
AST (TGO)	Enzima hepática (aspartato aminotransferase), presente em fígado, coração e músculos	U/L
Basófilos A	Contagem absoluta de basófilos	/µL
Basófilos R	Percentual de basófilos em relação aos leucócitos	%
Bastonetes A	Neutrófilos imaturos circulantes (contagem absoluta)	/µL
Bastonetes R	Percentual de neutrófilos em bastonetes	%
Bilirrubina Total	Produto da degradação da hemoglobina	mg/dL
Blastos A	Contagem absoluta de células blastocíticas	/µL
Blastos R	Percentual de blastos	%
Cálcio Ionizado	Fração biologicamente ativa do cálcio no sangue	mmol/L
Carboxihemoglobina	Hemoglobina ligada ao monóxido de carbono	%
CHCM	Concentração de hemoglobina corpuscular média	g/dL
Cloreto	Principal ânion extracelular; importante no equilíbrio ácido-base	mmol/L
CO_2 Total	Dióxido de carbono total (bicarbonato + CO_2 dissolvido)	mmol/L

Variável	Descrição	Unidade
Creatinina	Produto da degradação da creatina muscular; avalia função renal	mg/dL
CTO_2	Conteúdo total de oxigênio no sangue	mL/dL
Cálcio Ionizado	Fração biologicamente ativa do cálcio no sangue	mmol/L
DAV A	Desvio à esquerda absoluto (formas jovens de neutrófilos)	/µL
DAV R	Desvio à esquerda relativo	%
Eosinófilos A	Número absoluto de eosinófilos	/µL
Eosinófilos R	Percentual de eosinófilos entre os leucócitos	%
Excesso de Base	Quantidade de base necessária para normalizar o pH	mmol/L
Fósforo	Mineral envolvido no metabolismo energético e estrutura óssea	mg/dL
Glicose	Açúcar no sangue; principal fonte energética celular	mg/dL
Globulina	Fração proteica do plasma composta por imunoglobulinas e outras proteínas	g/dL
HCM	Hemoglobina corpuscular média	pg
HCO_3^-	Bicarbonato arterial; principal tampão do pH	mmol/L
Hemácias	Contagem total de glóbulos vermelhos	milhões/µL
Hemoglobina	Concentração total de hemoglobina no sangue	g/dL
Hemoglobina Reduzida	Hemoglobina não ligada ao oxigênio	%
Idade	Tempo de vida do paciente em anos completos	anos
KPTT SEG	Tempo de tromboplastina parcial ativada	s
Lactato	Produto do metabolismo anaeróbio; aumenta em hipóxia	mmol/L
Leucócitos	Contagem total de glóbulos brancos no sangue	/µL
Linfócitos A	Número absoluto de linfócitos	/µL
Linfócitos Atípicos A	Contagem absoluta de linfócitos com morfologia alterada	/µL
Linfócitos Atípicos R	Percentual de linfócitos atípicos	%
Linfócitos R	Percentual de linfócitos em relação aos leucócitos	%
Magnésio	Mineral intracelular envolvido em reações enzimáticas	mg/dL
Meta A	Contagem absoluta de metamielócitos	/µL

Variável	Descrição	Unidade
Meta R	Percentual de metamielócitos	%
Metahemoglobina (MetHb)	Forma oxidada da hemoglobina que não transporta oxigênio	%
Mielócitos A	Contagem absoluta de mielócitos circulantes	/µL
Mielócitos R	Percentual de mielócitos	%
Monócitos A	Número absoluto de monócitos	/µL
Monócitos R	Percentual de monócitos em relação aos leucócitos	%
NLCR	Razão neutrófilo-linfócito; marcador inflamatório	adimensional
Neutrófilos A	Número absoluto de neutrófilos	/µL
Neutrófilos R	Percentual de neutrófilos	%
Oxihemoglobina	Percentual de hemoglobina ligada ao oxigênio	%
P50	Pressão de O_2 para 50% de saturação da hemoglobina	mmHg
PCO_2	Pressão parcial de dióxido de carbono no sangue arterial	mmHg
PCR	Proteína C-reativa; marcador inflamatório agudo	mg/L
Plaquetas	Células envolvidas na coagulação sanguínea	mil/µL
PO_2	Pressão parcial de oxigênio no sangue arterial	mmHg
Potássio	Cátion intracelular essencial para função neuromuscular	mmol/L
Promielócitos A	Contagem absoluta de promielócitos	/µL
Promielócitos R	Percentual de promielócitos	%
Proteínas Totais	Soma da albumina e globulinas	g/dL
RDW	Amplitude de distribuição dos eritrócitos	%
Saturação (SaO_2)	Saturação de oxigênio da hemoglobina	%
Segmentados A	Número absoluto de neutrófilos segmentados	/µL
Segmentados R	Percentual de neutrófilos segmentados	%
Sódio	Principal cátion extracelular; regula osmolaridade	mmol/L
TAP	Tempo de atividade da protrombina	s
TAP RNI	Razão normalizada internacional da protrombina	adimensional
Ureia	Produto do metabolismo proteico excretado pelos rins	mg/dL
VCM	Volume corpuscular médio	fL
VG (Hematócrito)	Percentual do volume sanguíneo ocupado pelas hemácias	%

APÊNDICE C

Resultados da Seleção de variáveis via métodos sequenciais

Tabela 40 – Frequências de inclusão de variáveis por eliminação *backward*.

	$\alpha = 0.001$			$\alpha = 0.05$			$\alpha = 0.10$			$\alpha = 0.157$		
Variável	B(n)	B(m)	S(m)	B(n)	B(m)	S(m)	B(n)	B(m)	S(m)	B(n)	B(m)	S(m)
ASC	0.622	0.606	0.605	0.626	0.614	0.628	0.626	0.615	0.633	0.626	0.615	0.634
Hemoglobina	0.820	0.598	0.621	0.930	0.787	0.899	0.959	0.860	0.953	0.976	0.895	0.972
Bilirrubina Total	0.804	0.588	0.642	0.903	0.764	0.889	0.930	0.830	0.946	0.950	0.866	0.963
VCM	0.569	0.345	0.227	0.774	0.603	0.626	0.832	0.711	0.779	0.860	0.766	0.857
Idade	0.522	0.362	0.371	0.678	0.518	0.584	0.758	0.612	0.705	0.811	0.680	0.782
PCR	0.299	0.165	0.041	0.550	0.407	0.294	0.659	0.536	0.511	0.738	0.620	0.650
Monócitos R	0.387	0.262	0.214	0.530	0.412	0.414	0.596	0.508	0.528	0.643	0.564	0.597
NLCR	0.270	0.160	0.041	0.498	0.383	0.268	0.622	0.505	0.439	0.702	0.593	0.575
Eosinófilos R	0.293	0.209	0.108	0.503	0.385	0.292	0.622	0.496	0.436	0.701	0.576	0.559
Cálcio Ionizado	0.304	0.225	0.171	0.434	0.348	0.303	0.519	0.426	0.391	0.584	0.491	0.466
Lactato	0.298	0.230	0.236	0.391	0.319	0.308	0.460	0.383	0.368	0.522	0.444	0.423
Creatinina	0.335	0.289	0.278	0.402	0.363	0.334	0.448	0.411	0.373	0.494	0.459	0.407
pH	0.209	0.149	0.105	0.338	0.271	0.232	0.431	0.362	0.311	0.512	0.431	0.379
Ureia	0.201	0.136	0.078	0.330	0.255	0.191	0.408	0.333	0.290	0.473	0.394	0.364
HCM	0.143	0.108	0.027	0.332	0.290	0.086	0.486	0.435	0.187	0.601	0.550	0.310
Plaquetas	0.155	0.108	0.031	0.290	0.218	0.130	0.374	0.293	0.221	0.438	0.359	0.309
Glicose	0.121	0.082	0.010	0.289	0.218	0.085	0.405	0.332	0.188	0.497	0.423	0.298
Neutrófilos R	0.105	0.081	0.011	0.281	0.251	0.071	0.421	0.388	0.155	0.528	0.485	0.266
CHCM	0.085	0.071	0.008	0.232	0.217	0.045	0.358	0.329	0.125	0.463	0.421	0.229
p50	0.145	0.114	0.080	0.214	0.187	0.131	0.263	0.248	0.182	0.320	0.305	0.227
CO_2 Total	0.072	0.058	0.006	0.227	0.192	0.046	0.368	0.322	0.111	0.488	0.440	0.208
Monócitos A	0.129	0.098	0.039	0.228	0.211	0.101	0.298	0.293	0.156	0.355	0.356	0.206
Saturação	0.108	0.074	0.014	0.226	0.186	0.080	0.302	0.277	0.142	0.366	0.354	0.201
Segmentados R	0.073	0.062	0.004	0.235	0.219	0.039	0.366	0.341	0.104	0.472	0.439	0.199
Leucócitos	0.096	0.096	0.015	0.214	0.235	0.066	0.315	0.346	0.119	0.408	0.430	0.195
Segmentados A	0.081	0.082	0.013	0.186	0.207	0.056	0.272	0.302	0.105	0.351	0.380	0.165
DAV R	0.060	0.052	0.002	0.186	0.174	0.028	0.294	0.271	0.081	0.388	0.360	0.161
Sexo	0.056	0.049	0.002	0.161	0.145	0.029	0.264	0.224	0.076	0.353	0.305	0.143
Bastonetes A	0.069	0.077	0.012	0.165	0.189	0.045	0.251	0.282	0.083	0.323	0.356	0.140
RDW	0.074	0.064	0.014	0.148	0.130	0.044	0.212	0.199	0.085	0.283	0.266	0.132
pCO_2	0.055	0.058	0.004	0.146	0.164	0.031	0.224	0.261	0.075	0.306	0.344	0.125
Linfócitos A	0.063	0.056	0.006	0.154	0.164	0.038	0.232	0.261	0.075	0.299	0.326	0.122
Potássio	0.046	0.036	0.001	0.137	0.125	0.018	0.227	0.211	0.052	0.302	0.290	0.103
COHb	0.043	0.047	0.002	0.120	0.133	0.018	0.189	0.202	0.049	0.256	0.275	0.096
Magnésio	0.032	0.032	0.002	0.108	0.108	0.018	0.175	0.180	0.047	0.242	0.251	0.086
Sódio	0.035	0.031	0.000	0.113	0.125	0.013	0.182	0.204	0.040	0.253	0.282	0.080
Cloreto	0.032	0.031	0.002	0.112	0.121	0.012	0.183	0.197	0.039	0.256	0.268	0.075
pO_2	0.025	0.024	0.002	0.089	0.089	0.011	0.155	0.162	0.033	0.224	0.236	0.063
MetHb	0.015	0.023	0.001	0.054	0.071	0.004	0.102	0.129	0.012	0.158	0.187	0.032

Nota: B(n): *Bootstrap*(n); B(m): *Bootstrap*(m); S(m): *Subsampling*(m); R: Valores Relativos (%); A: Valores Absolutos (contagem celular); VCM: Volume Corpuscular Médio; PCR: Proteína C-Reativa; NLCR: Razão Neutrófilo-Linfócito; HCM: Hemoglobina Corpuscular Média; CHCM: Concentração de Hemoglobina Corpuscular Média; p50: Pressão de oxigênio a 50% de saturação da hemoglobina; CO_2: Dióxido de Carbono; RDW: Amplitude de distribuição dos Eritrócitos; pCO_2: Pressão Parcial de Dióxido de Carbono; DAV: Desvio absoluto reportado pelo sistema laboratorial; COHb: Carboxihemoglobina; PO_2: Pressão Parcial de Oxigênio; MetHb: Metahemoglobina.

Tabela 41 – Frequências de inclusão de variáveis por seleção *forward*.

	$\alpha = 0.001$			$\alpha = 0.05$			$\alpha = 0.10$			$\alpha = 0.157$		
Variável	B(n)	B(m)	S(m)	B(n)	B(m)	S(m)	B(n)	B(m)	S(m)	B(n)	B(m)	S(m)
ASC	0.627	0.609	0.610	0.635	0.622	0.630	0.635	0.625	0.637	0.634	0.624	0.640
Bilirrubina Total	0.753	0.505	0.529	0.904	0.744	0.834	0.948	0.824	0.929	0.956	0.872	0.968
Hemoglobina	0.735	0.498	0.501	0.883	0.714	0.812	0.937	0.809	0.918	0.965	0.863	0.958
Idade	0.374	0.245	0.222	0.582	0.419	0.405	0.684	0.526	0.549	0.762	0.605	0.664
Creatinina	0.528	0.425	0.539	0.597	0.478	0.589	0.605	0.512	0.614	0.633	0.543	0.640
PCR	0.216	0.097	0.020	0.507	0.323	0.209	0.662	0.495	0.426	0.753	0.587	0.597
Lactato	0.442	0.333	0.384	0.500	0.404	0.479	0.544	0.452	0.546	0.578	0.484	0.579
Monócitos R	0.322	0.235	0.192	0.453	0.355	0.345	0.535	0.439	0.445	0.620	0.500	0.532
Eosinófilos R	0.243	0.168	0.076	0.436	0.316	0.234	0.547	0.421	0.373	0.635	0.493	0.495
NLCR	0.176	0.075	0.011	0.409	0.270	0.159	0.546	0.409	0.334	0.639	0.500	0.483
VCM	0.234	0.141	0.063	0.419	0.304	0.250	0.538	0.416	0.376	0.631	0.493	0.481
HCM	0.345	0.212	0.219	0.429	0.325	0.390	0.489	0.392	0.444	0.549	0.446	0.461
Cálcio Ionizado	0.235	0.174	0.097	0.380	0.296	0.224	0.475	0.393	0.318	0.544	0.460	0.399
Plaquetas	0.140	0.093	0.021	0.279	0.206	0.119	0.384	0.292	0.223	0.470	0.368	0.335
p50	0.156	0.127	0.068	0.261	0.207	0.135	0.331	0.262	0.206	0.375	0.324	0.273
Ureia	0.105	0.081	0.032	0.220	0.171	0.081	0.329	0.251	0.155	0.411	0.332	0.249
Glicose	0.072	0.047	0.004	0.228	0.155	0.048	0.330	0.251	0.129	0.430	0.342	0.239
Monócitos A	0.135	0.080	0.035	0.207	0.155	0.120	0.261	0.221	0.174	0.299	0.255	0.213
Saturação	0.066	0.041	0.003	0.172	0.136	0.046	0.241	0.213	0.101	0.331	0.281	0.160
pH	0.051	0.033	0.008	0.145	0.095	0.047	0.200	0.143	0.086	0.285	0.224	0.152
Sexo	0.044	0.034	0.002	0.118	0.101	0.023	0.189	0.173	0.050	0.273	0.231	0.104
RDW	0.065	0.068	0.012	0.125	0.124	0.032	0.199	0.180	0.062	0.264	0.252	0.097
Magnésio	0.023	0.021	0.001	0.090	0.077	0.009	0.160	0.137	0.030	0.237	0.209	0.072
CO_2 Total	0.021	0.022	0.003	0.076	0.058	0.010	0.143	0.124	0.031	0.236	0.191	0.071
COHb	0.023	0.026	0.001	0.097	0.083	0.009	0.155	0.152	0.030	0.220	0.222	0.061
Neutrófilos R	0.016	0.012	0.001	0.060	0.043	0.006	0.121	0.107	0.020	0.185	0.162	0.054
pCO_2	0.013	0.011	0.000	0.051	0.037	0.004	0.099	0.090	0.018	0.162	0.141	0.042
Potássio	0.016	0.011	0.001	0.064	0.056	0.005	0.128	0.132	0.022	0.203	0.201	0.042
Linfócitos A	0.019	0.019	0.002	0.058	0.063	0.009	0.112	0.106	0.018	0.179	0.156	0.041
DAV R	0.016	0.022	0.002	0.053	0.060	0.007	0.089	0.096	0.018	0.136	0.142	0.038
Sódio	0.011	0.009	0.000	0.051	0.051	0.002	0.113	0.109	0.012	0.166	0.165	0.035
Leucócitos	0.012	0.013	0.000	0.041	0.041	0.007	0.067	0.068	0.020	0.105	0.097	0.034
Bastonetes A	0.009	0.013	0.001	0.046	0.050	0.003	0.091	0.089	0.014	0.129	0.135	0.031
Cloreto	0.013	0.010	0.000	0.055	0.050	0.004	0.091	0.091	0.013	0.150	0.147	0.031
MetHb	0.016	0.026	0.002	0.052	0.064	0.008	0.099	0.115	0.018	0.150	0.164	0.030
pO_2	0.014	0.012	0.000	0.050	0.049	0.003	0.105	0.100	0.012	0.173	0.167	0.029
Segmentados A	0.013	0.012	0.001	0.045	0.039	0.006	0.081	0.069	0.014	0.135	0.133	0.029
Segmentados R	0.015	0.011	0.001	0.039	0.037	0.005	0.075	0.071	0.014	0.135	0.116	0.027
CHCM	0.016	0.025	0.004	0.043	0.044	0.005	0.089	0.082	0.012	0.169	0.144	0.022

Nota: B(n): *Bootstrap*(n); B(m): *Bootstrap*(m); S(m): *Subsampling*(m); R: Valores Relativos (%); A: Valores Absolutos (contagem celular); PCR: Proteína C-Reativa; NLCR: Razão Neutrófilo-Linfócito; VCM: Volume Corpuscular Médio; HCM: Hemoglobina Corpuscular Média; p50: Pressão de oxigênio a 50% de saturação da hemoglobina; RDW: Amplitude de distribuição dos Eritrócitos; CO_2: Dióxido de Carbono; COHb: Carboxihemoglobina; MetHb: Metahemoglobina; pCO_2: Pressão Parcial de Dióxido de Carbono; DAV: Desvio absoluto reportado pelo sistema laboratorial; PO_2: Pressão Parcial de Oxigênio; CHCM: Concentração de Hemoglobina Corpuscular Média.

Tabela 42 – Frequências de inclusão de variáveis por seleção *stepwise*.

	$\alpha = 0.001$			$\alpha = 0.05$			$\alpha = 0.10$			$\alpha = 0.157$		
Variável	B(n)	B(m)	S(m)	B(n)	B(m)	S(m)	B(n)	B(m)	S(m)	B(n)	B(m)	S(m)
ASC	0.625	0.610	0.608	0.633	0.622	0.629	0.634	0.623	0.637	0.633	0.624	0.639
Bilirrubina Total	0.752	0.504	0.523	0.900	0.722	0.832	0.937	0.826	0.937	0.960	0.867	0.966
Hemoglobina	0.735	0.493	0.501	0.872	0.719	0.801	0.937	0.803	0.915	0.956	0.862	0.958
Idade	0.370	0.250	0.218	0.575	0.424	0.392	0.678	0.518	0.547	0.761	0.605	0.662
PCR	0.204	0.096	0.018	0.515	0.319	0.204	0.654	0.488	0.412	0.751	0.602	0.598
Creatinina	0.482	0.402	0.506	0.506	0.432	0.544	0.519	0.454	0.533	0.529	0.481	0.542
Monócitos R	0.342	0.224	0.173	0.459	0.337	0.342	0.538	0.415	0.437	0.609	0.488	0.538
Lactato	0.408	0.320	0.384	0.448	0.367	0.452	0.476	0.394	0.477	0.523	0.437	0.493
Eosinófilos R	0.241	0.167	0.086	0.429	0.311	0.224	0.557	0.422	0.374	0.641	0.504	0.489
VCM	0.245	0.139	0.069	0.419	0.296	0.246	0.523	0.412	0.366	0.618	0.492	0.477
NLCR	0.170	0.077	0.011	0.390	0.260	0.156	0.543	0.380	0.323	0.626	0.487	0.477
HCM	0.320	0.219	0.209	0.405	0.304	0.384	0.429	0.369	0.441	0.453	0.395	0.431
Cálcio Ionizado	0.229	0.178	0.101	0.370	0.299	0.211	0.477	0.375	0.314	0.538	0.438	0.400
Plaquetas	0.136	0.084	0.024	0.282	0.190	0.119	0.382	0.268	0.223	0.444	0.344	0.324
P50	0.153	0.127	0.068	0.254	0.205	0.135	0.302	0.262	0.214	0.350	0.308	0.260
Ureia	0.117	0.081	0.032	0.230	0.168	0.088	0.321	0.242	0.170	0.403	0.322	0.244
Glicose	0.074	0.043	0.004	0.212	0.161	0.052	0.335	0.251	0.129	0.433	0.343	0.239
Monócitos A	0.118	0.077	0.035	0.199	0.160	0.133	0.233	0.200	0.187	0.280	0.258	0.203
Saturação	0.061	0.043	0.003	0.163	0.134	0.048	0.245	0.213	0.098	0.309	0.287	0.164
pH	0.059	0.037	0.010	0.135	0.094	0.047	0.207	0.158	0.095	0.277	0.211	0.155
RDW	0.064	0.060	0.011	0.127	0.121	0.033	0.194	0.183	0.052	0.261	0.226	0.101
Sexo	0.042	0.028	0.003	0.115	0.098	0.027	0.185	0.164	0.060	0.266	0.225	0.101
Magnésio	0.025	0.019	0.000	0.085	0.081	0.008	0.159	0.140	0.029	0.233	0.211	0.065
COHb	0.026	0.027	0.002	0.087	0.086	0.010	0.156	0.143	0.025	0.227	0.223	0.056
CO_2 Total	0.019	0.018	0.004	0.073	0.056	0.010	0.134	0.104	0.027	0.225	0.179	0.054
Neutrófilos R	0.013	0.015	0.001	0.063	0.047	0.007	0.122	0.103	0.020	0.183	0.169	0.051
Potássio	0.014	0.014	0.000	0.068	0.051	0.006	0.124	0.115	0.020	0.199	0.185	0.047
PCO_2	0.012	0.011	0.000	0.049	0.048	0.005	0.100	0.082	0.019	0.151	0.135	0.046
Linfócitos A	0.020	0.019	0.001	0.055	0.051	0.009	0.113	0.100	0.019	0.166	0.158	0.041
Segmentados A	0.011	0.010	0.001	0.043	0.038	0.006	0.085	0.076	0.020	0.127	0.122	0.034
Cloreto	0.012	0.011	0.000	0.050	0.053	0.003	0.096	0.091	0.014	0.150	0.147	0.033
Bastonetes A	0.011	0.012	0.000	0.043	0.045	0.003	0.085	0.085	0.011	0.131	0.131	0.032
Sódio	0.011	0.009	0.000	0.050	0.054	0.002	0.109	0.105	0.012	0.167	0.158	0.031
PO_2	0.011	0.013	0.000	0.051	0.047	0.003	0.104	0.104	0.014	0.170	0.163	0.031
Leucócitos	0.012	0.008	0.001	0.036	0.035	0.006	0.063	0.063	0.016	0.099	0.102	0.031
MetHb	0.019	0.026	0.001	0.047	0.056	0.004	0.099	0.109	0.011	0.156	0.163	0.030
DAV R	0.018	0.020	0.003	0.051	0.053	0.005	0.088	0.091	0.014	0.142	0.133	0.028
Segmentados R	0.012	0.008	0.000	0.041	0.043	0.007	0.071	0.075	0.008	0.113	0.114	0.028
CHCM	0.015	0.021	0.004	0.031	0.037	0.005	0.067	0.067	0.008	0.131	0.117	0.015

Nota: B(n): *Bootstrap*(n); B(m): *Bootstrap*(m); S(m): *Subsampling*(m); R: Valores Relativos (%); A: Valores Absolutos (contagem celular); PCR: Proteína C-Reativa; VCM: Volume Corpuscular Médio; NLCR: Razão Neutrófilo-Linfócito; HCM: Hemoglobina Corpuscular Média; p50: Pressão de oxigênio a 50% de saturação da hemoglobina; RDW: Amplitude de distribuição dos Eritrócitos; COHb: Carboxihemoglobina; CO_2: Dióxido de Carbono; pCO_2: Pressão Parcial de Dióxido de Carbono; PO_2: Pressão Parcial de Oxigênio; MetHb: Metahemoglobina; DAV: Desvio absoluto reportado pelo sistema laboratorial; CHCM: Concentração de Hemoglobina Corpuscular Média.