UNIVERSIDADE FEDERAL DE PERNAMBUCO
CENTRO DE INFORMÁTICA
CURSO DE SISTEMAS DE INFORMAÇÃO

ALLYSON RYAN EMILIANO DA SILVA

ERICK DANIEL ALVES DE LIMA

LUCAS GABRIEL DE OLIVEIRA SILVA

MARIA EDUARDA DE LIMA GOMES

Aplicação do Metódo CRISP-DM para Diagnóstico Hospitalar Precoce de Sepse

Recife - PE

2025

ALLYSON RYAN EMILIANO DA SILVA

ERICK DANIEL ALVES DE LIMA

LUCAS GABRIEL DE OLIVEIRA SILVA

MARIA EDUARDA DE LIMA GOMES

Aplicação do Metódo CRISP-DM para Diagnóstico Hospitalar Precoce de Sepse

Experimentação no campo de dados, machine e deep learning envolvendo aspectos dos segmentos de engenharia, ciência e análise apresentado ao Curso de Sistemas de Informação do Centro de Informática da Universidade Federal de Pernambuco, como requisito parcial para avaliação da disciplina de Inteligência Artificial - IF1017.

Orientador(a): FLÁVIO ARTHUR OLIVEIRA DOS SANTOS

Recife - PE

2025

Aplicação do Metódo CRISP-DM para Diagnóstico Hospitalar Precoce de Sepse

ALLYSON RYAN EMILIANO DA SILVA

ERICK DANIEL ALVES DE LIMA

LUCAS GABRIEL DE OLIVEIRA SILVA

MARIA EDUARDA DE LIMA GOMES

Experimentação no campo de dados, machine e deep learning envolvendo aspectos dos segmentos de engenharia, ciência e análise apresentado ao Curso de Sistemas de Informação do Centro de Informática da Universidade Federal de Pernambuco, como requisito parcial para avaliação da disciplina de Inteligência Artificial - IF1017.

Aprovado em 15 de AGOSTO de 2025.

Prof.(a) FLÁVIO ARTHUR OLIVEIRA DOS SANTOS

Universidade Federal de Pernambuco

SUMÁRIO

1 METÓDO

1.1 Entendimento do problema

1.1.1 Contextualização do problema (Background)

A sepse, também conhecida como septicemia ou infecção generalizada, trata-se de uma condição médica alarmante e com alta reverberação em óbito, pode ser descrita como uma reação demasiada do organismo a uma infecção, normalmente bacteriana, impactando, inflamando tecidos diversos e desencadeando a fatal falência múltipla de órgãos após uma hipotensão arterial, nomeada como choque séptico (Drauzio Varella, 2025). Segundo o Centers for Disease Control and Prevention (CDC) (Centers for Disease Control and Prevention (CDC), 2024) no Estados Unidos a problemática é responsável por aproximadamente 200 mil mortes anualmente, resultantes das 1.7 milhões de vitimas, representando uma taxa de fatalidade consideravelmente. Esses números ocupam mais de um terço dos óbitos hospitalares no país. Escalando esses números para nivel global atinge-se cerca de 30 milhões de pessoas a cada ano, além das subnotificações em regiões mais carentes, as quais dentre essas falecem mais de 6 milhões, sendo uma parcela marjoritária crianças de diferentes faixas etárias.

1.1.2 Objetivos do negócio e critérios de sucesso

O custo financeiro torna-se óbvio indicando o desembolso acima dos 24 bilhões de dolláres para os norte-americanos, logo sendo a condição hospitalar mais custosa, ainda que o modelo padrão da nação envolva a privatização, há perdas financeiras por programas governamentais para populações vulneráveis e planos de saúde que fornecem apoio monetário abaixo do tratamento exigido para a condição. Esses gastos se dão por pacientes nesse cenário exigerem longas internações, monitoramento constante, medicamentos caros e normalmente aparelhos respiratórios conciliados a tratamentos renais intensivos. Contudo, há um agravante: maior parte desse custeamento é consequência de casos que são diagnosticados tardiamente, pós "janela de ouro", ou seja, na admissão hospitalar, o que pode ser precavido, pois facilmente há uma confirmação do caso examinando os níveis de lactato na corrente sanguínea. Esse fato é desastroso porquê influencia fortemente na alta taxa de mortalidade impactando a reputação dos hospitais. Isso é fruto de uma combinação perigosa: a sepse é camuflada inicialmente, sendo muito similar a viroses comuns, assim facilmente confudida diante de sintomas como febre, taquicardia, dificuldades respiratórias, pele pálida, todavia o avanço é extremamente rápido, dessa forma a observação correta ocorre apenas quando o paciente está em estado critico. Assim, é notório que há perdas financeiras para órgãos públicos ou de esfera privada, impactos em suas reputações diante da falha na saúde

coletiva demonstrada pela coleção de mortes, por isso é válido qualquer esforço que busque a mitigação. Dessa forma, o problema de negócio a ser abordado é a falta de diagnóstico precoce da sepse, que agrava a condição clínica, aumenta a mortalidade e eleva drasticamente os custos hospitalares e atinge a missão das organizações de saúde.

1.1.3 Inventário de recursos

Em primeiro aspecto é válido enfatizar que os dados trabalhados advém originalmente do seguinte dataset (Soni, 2022). A fim de validar e enriquecer os esforços envolvidos na imersão do problema, foram realizado pesquisas de outros campos divergentes e similares da tecnologia com o foco em tratar tal problemática, buscando melhor compreensão da mesma no cenário atual. Dentre essas é válido apresentar o DeepAise, modelo de rede neural recorrente que fornece pontuações de risco de sepse em tempo real, com AUC de até 0,90 e baixa taxa de falsos positivos (Araújo; Silva, 2023). De forma similar há o (Moor *et al.*, 2019), um estudo Dinamarquês que utilizou redes neurais profundas em dados de prontuários eletrônicos para prever sepse até 24 horas antes do início, com AUROC variando de 0,756 a 0,856. Por fim, novamente no cenário nacional destaca-se o "Robô Laura"(Costa, 2022), sistema brasileiro de IA que analisa sinais vitais e exames laboratoriais para emitir alertas precoces de sepse, auxiliando na tomada de decisões clínicas.

1.1.4 Objetivos do experimento e critérios de sucesso

Através da ciência de dados munida de algoritmos aplicados para mineração de dados visando alcançar a descoberta e exploração de padrões sútis, notados por esses por intermédio de associações, classes, categorizações, tendências, desvios, similaridades, caracterizações, discriminações e adjacentes almeja-se consolidar um modelo que preveja uma ocorrência de sepse com alta acurácia e sensibilidade, de forma que seja eficiente com relação ao tempo hábil para uma ação médica que interveia no tratamento, trazendo uma resposta média que precise o quadro em até 6 horas. Dessa forma, com uma integração desencadeando uma otimização no inicio do tratamento e redução dos custos, por evitar quadros mais graves, assim reduzindo a mortalidade hospitalar. Evidencia-se que a métrica protagonista trata-se da maior agilidade e preciosidade no diagnóstico hospitalar de pacientes com sepse.

1.1.5 Requisitos, suposições e restrições

Os percalços são apontados nas esferas distintas do segmento de dados. Na área de descoberta e engenharia de dados, encontra-se dificuldades referentes a qualidade e riqueza dos dados nos assets disponiveis enquanto visando-se a ciência de dados determina-se a curva de aprendizado na aplicação dos algoritmos e o reconhecimento

de requisitos que qualifique-os, além da pré exigência da engenharia no aspecto da fonte que esses modelos devem consumir e a baixa compreensão sobre o problema tratado. Por fim, visualiza-se ainda uma acessibilidade no aspecto que possam ser realmente útil, ainda que exija-se melhorias, para o campo hospitalar, sendo interpretável no contexto das pessoas envolvidas de forma que seja possivel uma análise dos dados para insights ágeis.

1.1.6 Riscos e contigências

Além da curva de aprendizado referente aos requisitos multidisciplinares, a experimentação envolve riscos comuns a outros casos de manipulação e análise dos dados, dentre eles está a interpretação incorreta dos dados clínicos, o que pode acarretar a modelos enviesados ou imprecisos. No aspecto da qualidade dos recursos disponiveis ressalta-se que a generalização de um modelo preditivo treinado em determinada base de dados pode não funcionar adequadamente em hospitais com diferentes perfis populacionais, protocolos clínicos ou tecnologias de coleta de dados. Contudo, está disposto uma gama de profissionais ativos da área que estão dispostos a contribuir para o avanço mediante interesse e contato.

1.1.7 Custos e benefícios potenciais

O investimento está vinculado à principalmente tempo-esforço dos envolvidos, dado que a infraestrutura computacional exigida não é HPC (High Power Computing), logo facilmente disponivel no ambiente público da universidade. Logo, os beneficios sobrepõem-se esses custos indiretos. Um modelo eficaz de predição precoce de sepse é de tamanha contribuição para literatura de tecnologia e saúde, pois funcional em implantação reduz significativamente o tempo de resposta clínica, diminui a gravidade dos casos atendidos, amenizando internações em UTI, baixando os custos com antibióticos e suporte intensivo, além de preservar vidas humanas e reputações das entidades adotantes.

1.1.8 Plano preliminar do projeto

O experimento segue a metodologia CRISP-DM, adotando suas fases para guiar o desenvolvimento do projeto com foco na detecção precoce de sepse. A seguir, estão descritas as etapas conforme o modelo:

Fase 2 – Entendimento dos dados: Realiza-se uma análise exploratória aprofundada dos dados disponibilizados, com verificação da qualidade, formato, consistência e conformidade. São aplicadas estatísticas descritivas e técnicas de visualização para detectar padrões relevantes, anomalias e outliers, apoiando o diagnóstico precoce e a tomada de decisão sobre o pré-processamento.

Fase 3 – Preparação dos dados: Inclui as etapas de limpeza, integração e transformação dos dados, além do tratamento de valores ausentes. Realiza-se a normalização das variáveis e a seleção de atributos com maior poder preditivo, a fim de otimizar o desempenho dos modelos nas etapas seguintes.

Fase 4 – Modelagem: Aplicação de algoritmos supervisionados como Random Forest, Gradient Boosting e redes neurais artificiais. Cada modelo é avaliado com validação cruzada e ajuste de hiperparâmetros (tuning), visando maximizar métricas de desempenho em cenários desbalanceados.

Fase 5 – Avaliação: Os modelos gerados são comparados com base em métricas como AUC, precisão, recall e F1-score, com ênfase na *sensibilidade* (recall) para garantir uma alta taxa de detecção precoce de sepse. A escolha final será baseada não apenas na acurácia geral, mas no equilíbrio entre falso-positivos e falso-negativos.

Fase 6 – Implementação: Estuda-se a viabilidade de integração do modelo em sistemas hospitalares, como prontuários eletrônicos, dashboards interativos ou sistemas de alerta clínico. A implementação será discutida com foco na aplicabilidade em ambiente real e no impacto clínico.

Estimativa preliminar de tempo e esforço: A execução do projeto está estimada entre **6 a 10 semanas**, com marcos de revisão a cada fase para acompanhamento do progresso. Considera-se o envolvimento de um a dois pesquisadores em tempo parcial, com esforço concentrado especialmente nas fases de preparação e modelagem. A documentação será mantida de forma contínua, permitindo ajustes conforme necessidade e garantindo rastreabilidade das decisões tomadas.

A estimativa inicial de execução é de 6 a 10 semanas, com marcos de revisão a cada fase. O projeto será documentado continuamente e avaliado com base nos critérios de sucesso definidos com foco no impacto clínico.

1.1.9 Importância da compreensão do problema

A compreensão detalhada do problema é uma etapa metodológica habitualmente ignorada, isso é um erro considerável no desenvolvimento, pois essa fase deve ser contínua e é basal para qualquer outra, sendo crucial na entrega de valor. Sem um entendimento aprofundado do contexto clínico da sepse, de seus impactos e de como o sistema de saúde responde a ela, cria-se o risco de produzir modelos irrelevantes, inviáveis ou até perigosos. Compreender o âmbito contextual do negócio em aspectos operacionais e econômicos permite direcionar esforços num escopo valoroso. Assim, o sucesso do projeto possui dependência da clareza e profundidade da fase de entendimento do problema, sendo este o alicerce que assegura a construção de soluções realmente transformadoras para a saúde pública.

1.2 Entendimento dos dados

Em primeiro momento, destaca-se que muitas das recomendações advém de (Kore, 2023) e (Pabba, 2023).

1.2.1 Coleta Inicial dos Dados

Originalmente, houve uma dificuldade significativa no que se diz respeito da identificação dos pacientes dado que a coluna "ID"foi mutilada do dataset, porém com algumas buscas encontrou-se trabalhos que obtiveram acesso privilegiado mantendo a coluna para o mesmo asset, dessa forma houve enxertamento de tal coluna através da comparação com o dataset disponibilizado em (PhysioNet, 2019) , logo indo além da referência utilizada (Soni, 2022) .

1.2.2 Descrição Geral da Base de Dados

Inicialmente, a base de dados fornecida apresenta 1.552.210 registros, possuindo 42 colunas, sendo apenas uma delas apontada como target, tal é a indicativa de sepse, consistindo-se como uma variável binária, dictômica ("Ou possui sepse ou não possui"). Tais registros dão-se como uma série temporal que apresenta o acompanhamento de pacientes na unidade hospitalar, ou seja, esse 1,5> milhões de registros referem-se a uma quantia menor de pacientes, englobando cada um numa quantidade diversa que relaciona-se com o tempo que tal foi acompanhado, sendo cada linha atualizações do monitoramento com possiveis novas informações a serem registradas nas colunas. Um paciente qualquer pode estar vinculado a várias linhas, sendo cada uma delas referenciada como um momento em hora, ou seja, informações especificas do paciente no momento-hora 1, no momento-hora 2 e assim por conseguinte, não tratando-se de um timestamp, mas de uma variável quantitativa em inteiro (discreta) e não aditiva. A inserção de IDs ocorreu a fim de que houvesse otimização de tempo, sem a necessidade de estipular com margens de erro valores de identificação derivados dos comportamentos da coluna "Hora", o qual apresenta uma sequência numérica que é reiniciada demarcando os registros de um novo paciente, trazendo assim limites. Dessa forma, através de uma coleta de valores distintos conclui-se que há cerca de 40.336.

1.2.3 Verificação da Qualidade de Dados

Ainda a respeito das qualidades das colunas a Tabela 1 informa a categorização respectiva de cada atributo, observando-se que há possibilidades de explorar agregações por faixas pré-definidas posteriormente, além disso há uma quantia majoritária de variáveis continuas, dado que o contexto trata-se de medidas corporais. É válido ressaltar

que tais colunas não estão originalmente apresentadas com tais nomes tendo sido renomeadas conforme as relações indicadas abaixo.

Tabela 1 – Variáveis renomeadas e classificações

Nome Original	**Renome**	**Tipo de Variável**	**Classificação**
Hour	Hora	Quantitativa	Discreta
HR	FrequênciaCardíaca	Quantitativa	Contínua
O2Sat	SaturaçãoDeOxigênio	Quantitativa	Contínua
Temp	Temperatura	Quantitativa	Contínua
SBP	PressãoArterialSistólica	Quantitativa	Contínua
MAP	PressãoArterialMédia	Quantitativa	Contínua
DBP	PressãoArterialDiastólica	Quantitativa	Contínua
Resp	FrequênciaRespiratória	Quantitativa	Contínua
EtCO2	CO2NoFimDaExpiração	Quantitativa	Contínua
BaseExcess	ExcessoDeBase	Quantitativa	Contínua
HCO3	Bicarbonato	Quantitativa	Contínua
FiO2	FraçãoDeOxigênioInspirado	Quantitativa	Contínua
pH	pH	Quantitativa	Contínua
PaCO2	PressãoParcialDeCO2	Quantitativa	Contínua
SaO2	SaturaçãoArterialDeOxigênio	Quantitativa	Contínua
AST	AspartatoAminotransferase	Quantitativa	Contínua
BUN	NitrogênioUreicoNoSangue	Quantitativa	Contínua
Alkalinephos	FosfataseAlcalina	Quantitativa	Contínua
Calcium	NívelDeCálcio	Quantitativa	Contínua
Chloride	NívelDeCloreto	Quantitativa	Contínua
Creatinine	NívelDeCreatinina	Quantitativa	Contínua
Bilirubin_direct	BilirrubinaDireta	Quantitativa	Contínua
Glucose	NívelDeGlicose	Quantitativa	Contínua
Lactate	NívelDeLactato	Quantitativa	Contínua
Magnesium	NívelDeMagnésio	Quantitativa	Contínua
Phosphate	NívelDeFosfato	Quantitativa	Contínua
Potassium	NívelDePotássio	Quantitativa	Contínua
Bilirubin_total	BilirrubinaTotal	Quantitativa	Contínua
TroponinI	TroponinaI	Quantitativa	Contínua
Hct	Hematócrito	Quantitativa	Contínua
Hgb	Hemoglobina	Quantitativa	Contínua
PTT	TempoDeTromboplastinaParcial	Quantitativa	Contínua
WBC	ContagemDeLeucócitos	Quantitativa	Contínua
Fibrinogen	NívelDeFibrinogênio	Quantitativa	Contínua
Platelets	ContagemDePlaquetas	Quantitativa	Contínua
Age	Idade	Quantitativa	Contínua
Gender	Sexo	Qualitativa	Binária
Unit1	UnidadeMICU	Qualitativa	Binária
Unit2	UnidadeSICU	Qualitativa	Binária
HospAdmTime	TempoDesdeAdmissãoHospitalar	Quantitativa	Contínua
ICULOS	TempoDeInternaçãoNaUTI	Quantitativa	Contínua
SepsisLabel	IndicadorDeSepse	Qualitativa	Binária
Patient_ID	ID	Qualitativa	Nominal

Por fim, apresenta-se a Tabela 2 contendo informações alarmantes, tratando-se da taxa de povoamento de cada atributo, a qual apenas 15/43 deles tem um valor de povoamento acima 60%, estando a maioria abaixo de 10%. Normalmente, na exploração dos dados considera-se uma exigência padrão o limite de 70% de preenchimento, contudo tal expectativa reduziria ainda mais o valor do dataset, limitando-o a 11 colunas. Diante disso, foi imposta uma tolerância de 60%, mantendo para trabalho 15 colunas, considerando-as proveitosas, mas tendo como perspectiva uma estratégia de imputação pela constante ou repetição do dado mais recente faz-se interessante o trabalho também com a coluna 'temperatura', apesar de haver apenas 33,84% de população. Assim, decorrerá-se em 16 colunas. Foi versionado o dataset original para que houvesse uma filtragem apenas dos pacientes que foram diagnósticados com a condição, contudo a taxa de valores nulos prosseguiu extremamente próxima (Tabela 3), inclusive apresentando aumento em algumas caracteristicas, logo é notório que não há nenhuma causualidade nos poucos registros dessas colunas e o aparecimento de sepse. Indo além, quando trata-se de duplicidade tal é esperada dentro do contexto para quaisquer colunas, sendo para análise mais relevante uma aproximação de valores especificos, ou seja, se há um gritante enquadramento, uma concentração numa faixa especifica, deve-se investigar se tal intervalo possui uma relação com alguma outra variável, estabelecendo causualidade se a hipótese for consolidada. Ainda é válido ressaltar que não há registros de duplicidade no que refere-se a agrupamentos entre hora e ID, além de não haver rows idênticas, tais retornos sinalizam boa coerência. IDs também não apresentam, mesmo possuindo mais de uma aparição em variação de hora, campos como idade e sexo distoantes.

Referente a tal coluna target, é nitido que há uma proporção baixa de pacientes que apresentam o quadro, ocupando apenas 7.27% (2.932) dos registros, podendo ser um fator determinante ou mais um desafio.

Tabela 2 – Povoamento dos campos

Coluna	**Não Nulos**	**% Não Nulos**
Hora	1552210	100.00
FrequênciaCardíaca	1398811	90.12
SaturaçãoDeOxigênio	1349474	86.94
Temperatura	525226	33.84
PressãoArterialSistólica	1325945	85.42
PressãoArterialMédia	1358940	87.55
PressãoArterialDiastólica	1065656	68.65
FrequênciaRespiratória	1313875	84.65
CO2NoFimDaExpiração	57636	3.71
ExcessoDeBase	84145	5.42
Bicarbonato	65028	4.19
FraçãoDeOxigênioInspirado	129365	8.33
pH	107573	6.93
PressãoParcialDeCO2	86301	5.56
SaturaçãoArterialDeOxigênio	53561	3.45
AspartatoAminotransferase	25183	1.62
NitrogênioUreicoNoSangue	106568	6.87
FosfataseAlcalina	24941	1.61
NívelDeCálcio	91331	5.88
NívelDeCloreto	70466	4.54
NívelDeCreatinina	94616	6.10
BilirrubinaDireta	2990	0.19
NívelDeGlicose	265516	17.11
NívelDeLactato	41446	2.67
NívelDeMagnésio	97951	6.31
NívelDeFosfato	62301	4.01
NívelDePotássio	144525	9.31
BilirrubinaTotal	23141	1.49
TroponinaI	14781	0.95
Hematócrito	137433	8.85
Hemoglobina	114591	7.38
TempoDeTromboplastinaParcial	45699	2.94
ContagemDeLeucócitos	99447	6.41
NívelDeFibrinogênio	10242	0.66
ContagemDePlaquetas	92209	5.94
Idade	1552210	100.00
Sexo	1552210	100.00
UnidadeMICU	940250	60.57
UnidadeSICU	940250	60.57
TempoDesdeAdmissãoHospitalar	1552202	100.00
TempoDeInternaçãoNaUTI	1552210	100.00
IndicadorDeSepse	1552210	100.00
ID	1552210	100.00

Tabela 3 – Povoamento dos campos para pacientes com sepse

Colunas	**Não Nulos**	**% Não Nulos**
Hora	27916	100.00
FrequênciaCardíaca	25699	92.06
SaturaçãoDeOxigênio	25024	89.64
Temperatura	9476	33.94
PressãoArterialSistólica	23674	84.80
PressãoArterialMédia	25265	90.50
PressãoArterialDiastólica	19573	70.11
FrequênciaRespiratória	23912	85.66
CO2NoFimDaExpiração	2997	10.74
ExcessoDeBase	2856	10.23
Bicarbonato	1726	6.18
FraçãoDeOxigênioInspirado	5020	17.98
pH	3730	13.36
PressãoParcialDeCO2	3155	11.30
SaturaçãoArterialDeOxigênio	1811	6.49
AspartatoAminotransferase	904	3.24
NitrogênioUreicoNoSangue	2528	9.06
FosfataseAlcalina	900	3.22
NívelDeCálcio	2404	8.61
NívelDeCloreto	1912	6.85
NívelDeCreatinina	2162	7.74
BilirrubinaDireta	130	0.47
NívelDeGlicose	5091	18.24
NívelDeLactato	1961	7.02
NívelDeMagnésio	2394	8.58
NívelDeFosfato	1645	5.89
NívelDePotássio	3681	13.19
BilirrubinaTotal	805	2.88
TroponinaI	320	1.15
Hematócrito	3133	11.22
Hemoglobina	2685	9.62
TempoDeTromboplastinaParcial	1224	4.38
ContagemDeLeucócitos	2325	8.33
NívelDeFibrinogênio	330	1.18
ContagemDePlaquetas	2099	7.52
Idade	27916	100.00
Sexo	27916	100.00
UnidadeMICU	15333	54.93
UnidadeSICU	15333	54.93
TempoDesdeAdmissãoHospitalar	27916	100.00
TempoDeInternaçãoNaUTI	27916	100.00
IndicadorDeSepse	27916	100.00
ID	27916	100.00

1.2.4 Exploração dos Dados

Os dados nulos apresentam correlação forte entre algumas colunas, como na Figura 1. Isso se dá por ser dados derivados ou de uma única medição, é notório isso no mapa de calor que expõe um quadrante entre as medidas relacionadas. Também tratando-se de campos contrários há uma forte relação de nulidade, mas tal não deve apresentar estranheza, conclui-se que quando pacientes não são registrados na UTI (MICU), há grande possibilidade de também não ter sido registrado na enfermagem (SICU), resumidamente não sendo amparado por tal triagem.

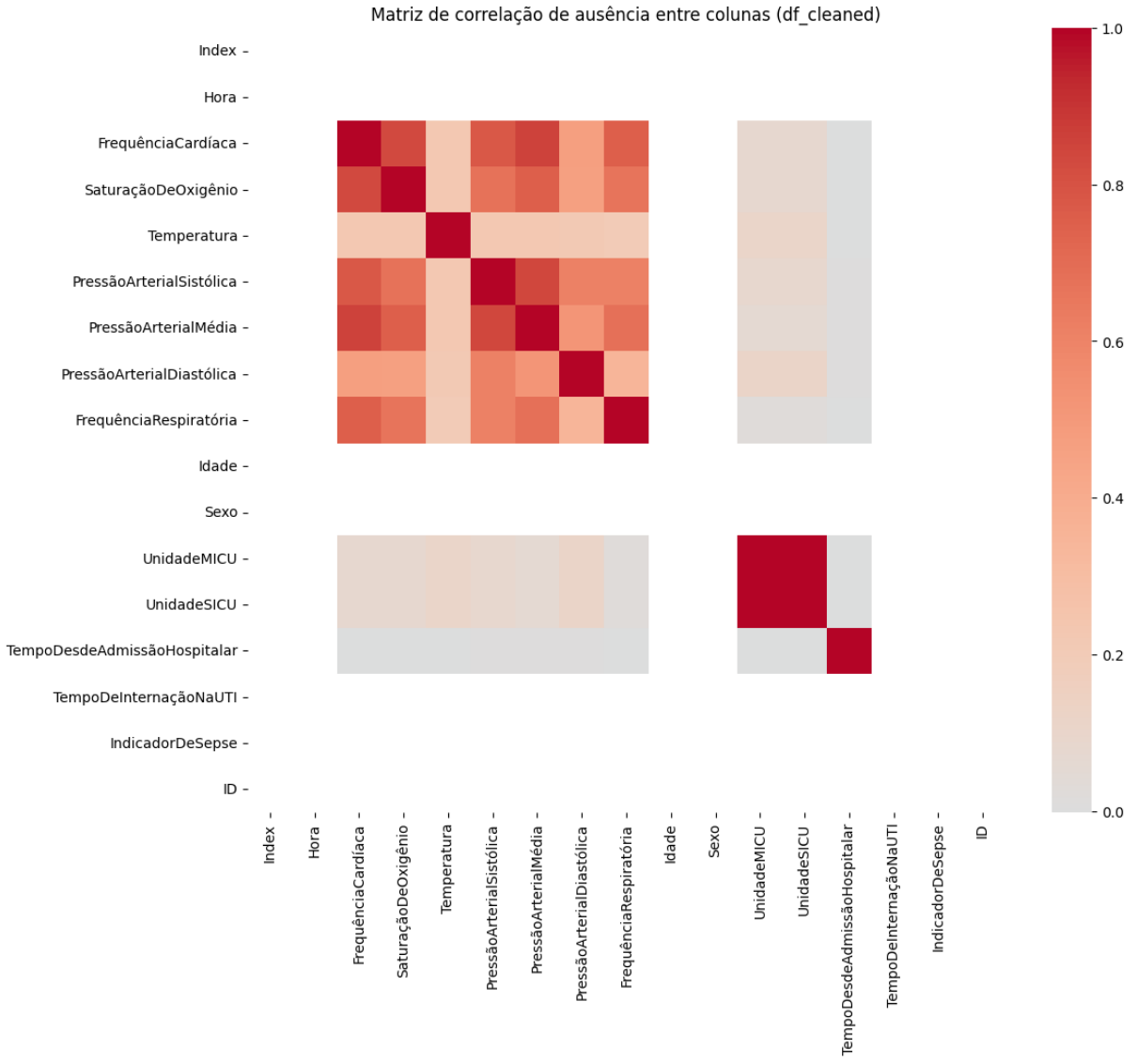

Figura 1 – Mapa de calor medidor da relação de nulidade entre campos.

Prosseguindo numa análise coerente, focando nas colunas que apresentam massa de valores significativas, antecipadamente foi observado a formatação de tais dados, a fim de visualizar possiveis correções. Observou-se que há uma qualidade considerável dos dados, com nenhuma coluna possuindo tipagens mistas, com taxas de distinção corretas em variáveis binárias (Ex: Sexo possui apenas dois valores distintos, assim

como IndicadorDeSepse) e unidades de medida aparentemente uniformes. Foram coletadas algumas medidas de dispersão e distribuição. A estatistica sumária foi realizada tendo a cautela se a medida é aditiva ou não, tendo-se como conclusão que não há medidas interessantemente e coerentemente aditivas, pois referem-se há indicadores continuos que registram momentos pontuais de cada paciente, sendo mais pertinente a realização de médias, máximos, mínimos, modas, variação e entre outras medições por paciente. Além disso, houve uma análise sobre os campos potenciais alvos de binning, o que pode auxiliar o modelo (Tabela 4).

Tabela 4 – Funções de agregação recomendadas por paciente e potencial para aplicação de faixas

Coluna	**Agregações Recomendadas**	**Aplicação de Faixas**
Hora	`min`, `max`, `count`	não
FrequênciaCardíaca	`mean`, `min`, `max`, `std`, `median`, `last`	Bradicardia, normal, taquicardia
SaturaçãoDeOxigênio	`mean`, `min`, `max`, `median`, `last`	<90%, 90–94%, <=95%
Temperatura	`mean`, `min`, `max`, `std`, `median`	Hipotermia, normal, febre, hipertermia
PressãoArterialSistólica	`mean`, `min`, `max`, `std`, `median`	Hipotensão, normal, hipertensão
PressãoArterialMédia	`mean`, `min`, `max`, `std`	Faixas para perfusão sistêmica
PressãoArterialDiastólica	`mean`, `min`, `max`, `std`	Faixas similares à sistólica
FrequênciaRespiratória	`mean`, `min`, `max`, `std`, `median`	Bradipneia, normal, taquipneia
Idade	—	Faixas etárias
Sexo	—	não
UnidadeMICU	`max`, `any`	não
UnidadeSICU	`max`, `any`	não
TempoDesdeAdmissãoHospitalar	`min`, `max`, `range`	Faixas temporais
TempoDeInternaçãoNaUTI	`max`	Curto, médio, longo prazo
IndicadorDeSepse	`max`, `any`, `sum`	não
ID	—	não

Ainda a respeito da estatísitica sumária (descritiva) desconsiderando colunas irrelevantes para análise (identificadores e variáveis binárias), mantem-se no escopo as variáveis numéricas contínuas de interesse. A Tabela 5 apresenta as principais estatísticas.

Tabela 5 – Estatísticas descritivas das variáveis numéricas

Coluna	**Desvio Padrão**	**Variância**	**Amplitude**
Frequência Cardíaca	17.33	300.16	260.00
Saturação de Oxigênio	2.94	8.63	80.00
Temperatura	0.77	0.59	29.10
Pressão Arterial Sistólica	23.23	539.71	280.00
Pressão Arterial Média	16.34	267.05	280.00
Pressão Arterial Diastólica	13.96	194.77	280.00
Frequência Respiratória	5.10	25.99	99.00
Idade	16.39	268.51	86.00
Tempo desde admissão hospitalar	162.26	26327.30	5390.85
Tempo de internação na UTI	29.01	841.31	335.00

A partir de tais estatisticas foram considerados os outliers diante do cumprimento do critério estabelecido pela regra do intervalo interquartil (IQR) para definição de outlier. Valores fora do intervalo $[Q1 - 1{,}5 \times IQR,\ Q3 + 1{,}5 \times IQR]$ foram classificados como outliers. A Tabela 6 resume os resultados.

Tabela 6 – Resumo de Outliers por IQR

Variável	**Q1**	**Q3**	**IQR**	**Lim. Inf.**	**Lim. Sup.**	**% Outliers**
Frequência Cardíaca	72.00	95.50	23.50	36.75	130.75	0.90%
Saturação de Oxigênio	96.00	99.50	3.50	90.75	104.75	1.60%
Temperatura	36.50	37.50	1.00	35.00	39.00	0.42%
Pressão Art. Sistólica	107.00	138.00	31.00	60.50	184.50	1.02%
Pressão Art. Média	71.00	92.00	21.00	39.50	123.50	1.41%
Pressão Art. Diastólica	54.00	72.00	18.00	27.00	99.00	1.05%
Frequência Respiratória	15.00	21.50	6.50	5.25	31.25	1.79%
Idade	51.68	74.00	22.32	18.20	107.48	0.15%
Tempo Admissão Hosp.	–47.05	–0.04	47.01	–117.56	70.47	13.73%
Tempo na UTI	11.00	34.00	23.00	–23.50	68.50	4.60%
Indicador de Sepse	0.00	0.00	0.00	0.00	0.00	1.80%
Index / Hora	9.00	33.00	24.00	–27.00	69.00	4.41%

Outra grande contribuição para compreensão dos dados refere-se a uma matriz de correlações (Figura 2) que aponta uma análise de bivariância visual. Na matriz apresentada pode-se verificar que TempoDeInternaçãoNaUTI é igual à coluna Hora, além disso as pressões têm uma alta correlação entre si, algo previsivel dado que são dados derivados e também coletados a partir da mesma fonte. A frequência respiratória parece também acompanhar as pressões arteriais. Por fim, observa-se as unidades MICU e SICU possuem correlação inversa, algo lógico dado que o paciente ocupa ou a enfermaria ou a UTI. Contudo, compactando tal matriz para focar na coluna target (IndicadorDeSepse), é ainda mais notório que não há correlação significativa com nenhuma outra variável conforme o indicador da Figura 3

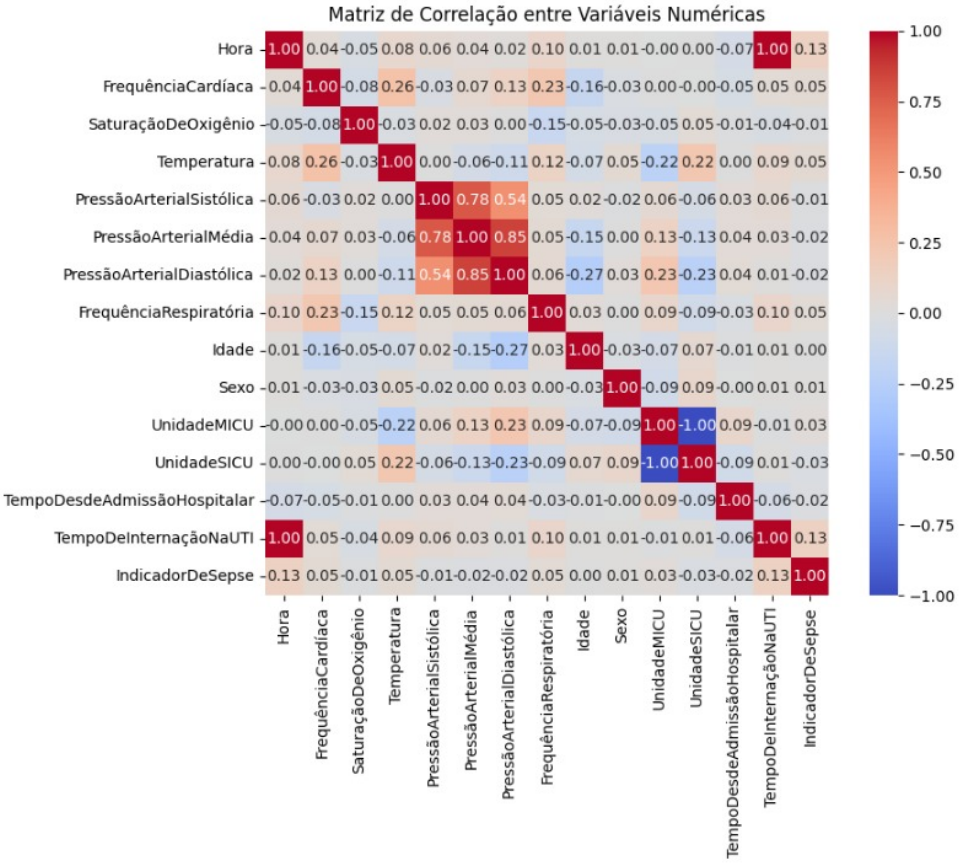

Figura 2 – Matriz de correlação entre variáveis númericas

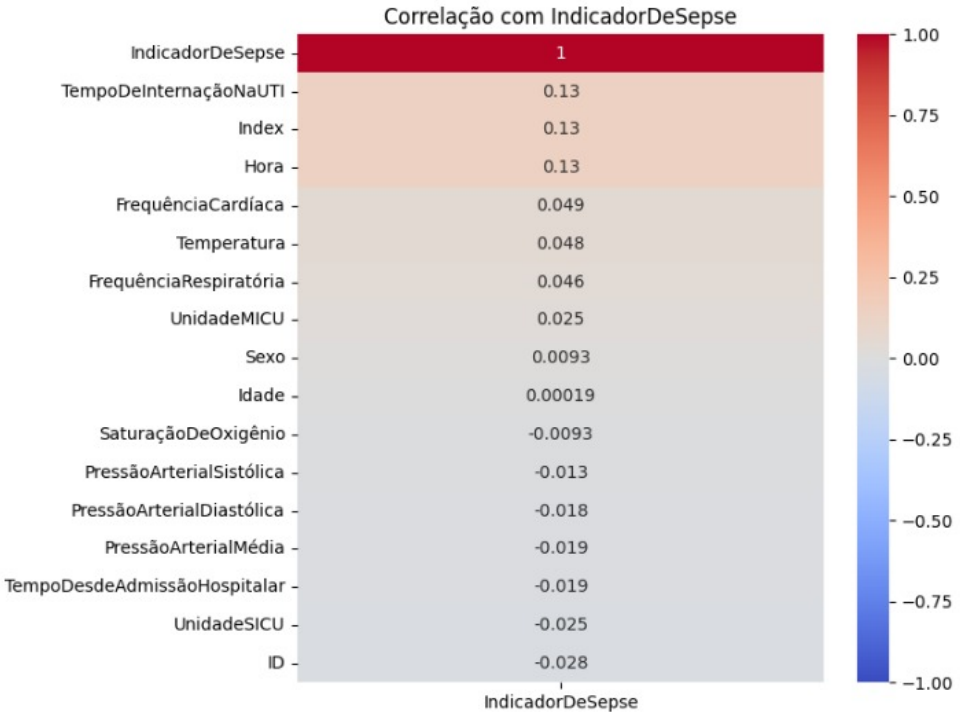

Figura 3 – Correlação com IndicadorDeSepse

1.2.4.1 Visualização dos Dados

Antecipadamente, é crucial compreender que a visualização de dados é essencial para facilitar a leitura da qualidade do dataset, contudo deve haver uma cautela na interpretação desses gráficos são alvos de diversos fatores númericos. Diante isso, faz-se necessário ir além dos dados nas estatísticas descritivas. Inicia-se focando na variável alvo (Indicador de Sepse) Na Figura 4 há um gráfico de barras (Count Plot)

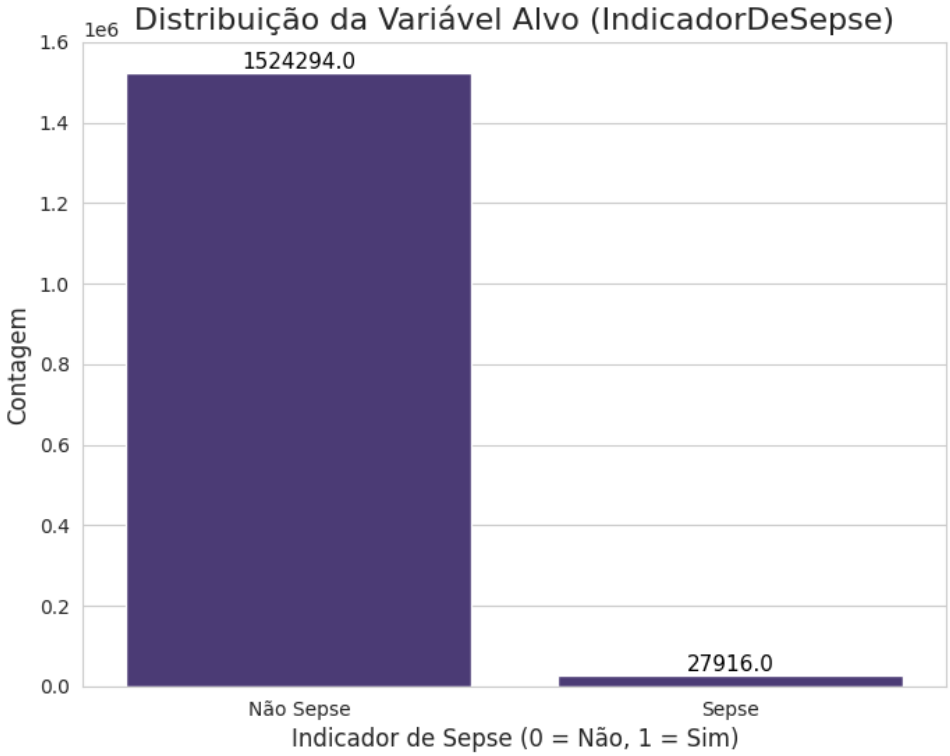

Figura 4 – Distribuição entre pacientes positivos e não positivos

a partir da contagem de 0s (Não Sepse) e 1s (Sepse) na coluna IndicadorDeSepse. O intuito é verificar o balanceamento das classes, tornando-se nitido que há pouquissimos registros de pacientes com sepse. Evidencia-se novamente o tamanho desse desequilíbrio, o que é fundamental para as próximas etapas de modelagem.

Na Figura 5 é possível observar a distribuição das principais colunas em histogramas, possuindo comportamentos já esperados ao encaixar-se em janelas previsiveis conforme os contextos que respeitam a faixa fisiológica. Há uma forte distribuição normal com assimétrias coerentes com os atributos analisados. Além disso, a partir deles já é notória a baixa presença de outliers. Reforça-se também que possuem múltiplos picos, isso se dá fortemente pelo contexto do atributo, por exemplo a temperatura comum humana é uma faixa curta entre 34,5º e 36º graus Celsius.

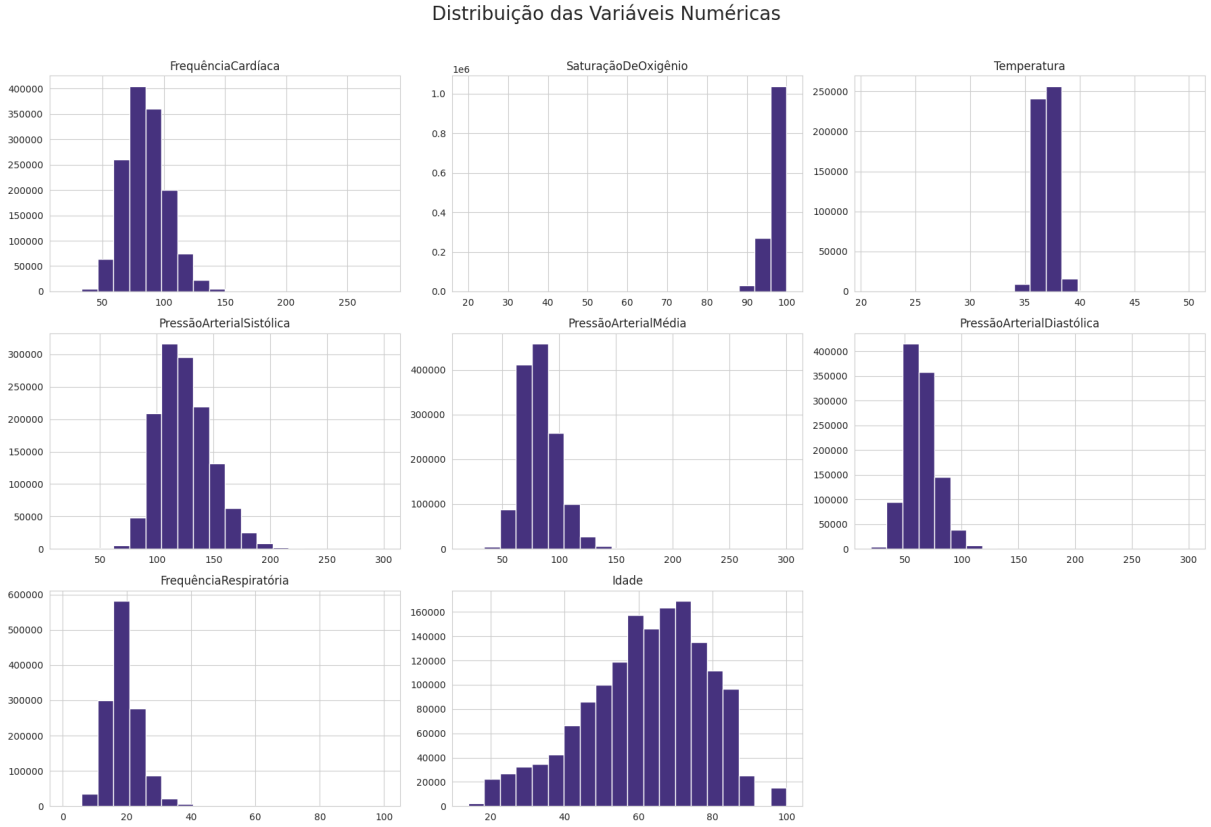

Figura 5 – Distribuição Univariada das Colunas

Contudo, ao observar boxplots (Figura 6) pode-se ter conclusões precipitadas e conflitantes com os histogramas, isso se dá pelo comportamento do gráfico em amostras massivas, tendo em perspectiva que ainda que proporcionalemente hajam poucos outliers em cada coluna, há uma pontuação gráfica para cada um, logo ainda que assumam um número inferior à 1% em relação ao total dado que são números grandes essa porcentagens trará poluição no boxplot. Outro fator que afeta o boxplot é o contexto fisiológico e a baixa quantia de pacientes que testam positivo, consequentemente o gráfico de não sepse apresentará baixos números em todos aspectos.

Traçando algumas colunas relevantes e posicionadas adequadamente na Figura 3, avalia-se e visualiza-se a evolução de sinais vitais de forma comparativa, buscando identificar algo sintomático, usa-se para essa evolução temporal o tempo de internação na UTI (abscissa). Pode-se elaborar isso através da comparação entre os padrões das Figuras 7 e 8. Entretanto, obter essa comparação a partir de uma única amostra totalmente sem critério é extremamente ineficiente, por isso separa-se em tendências em séries temporais sobre os atributos relevantes como na Figura 9.

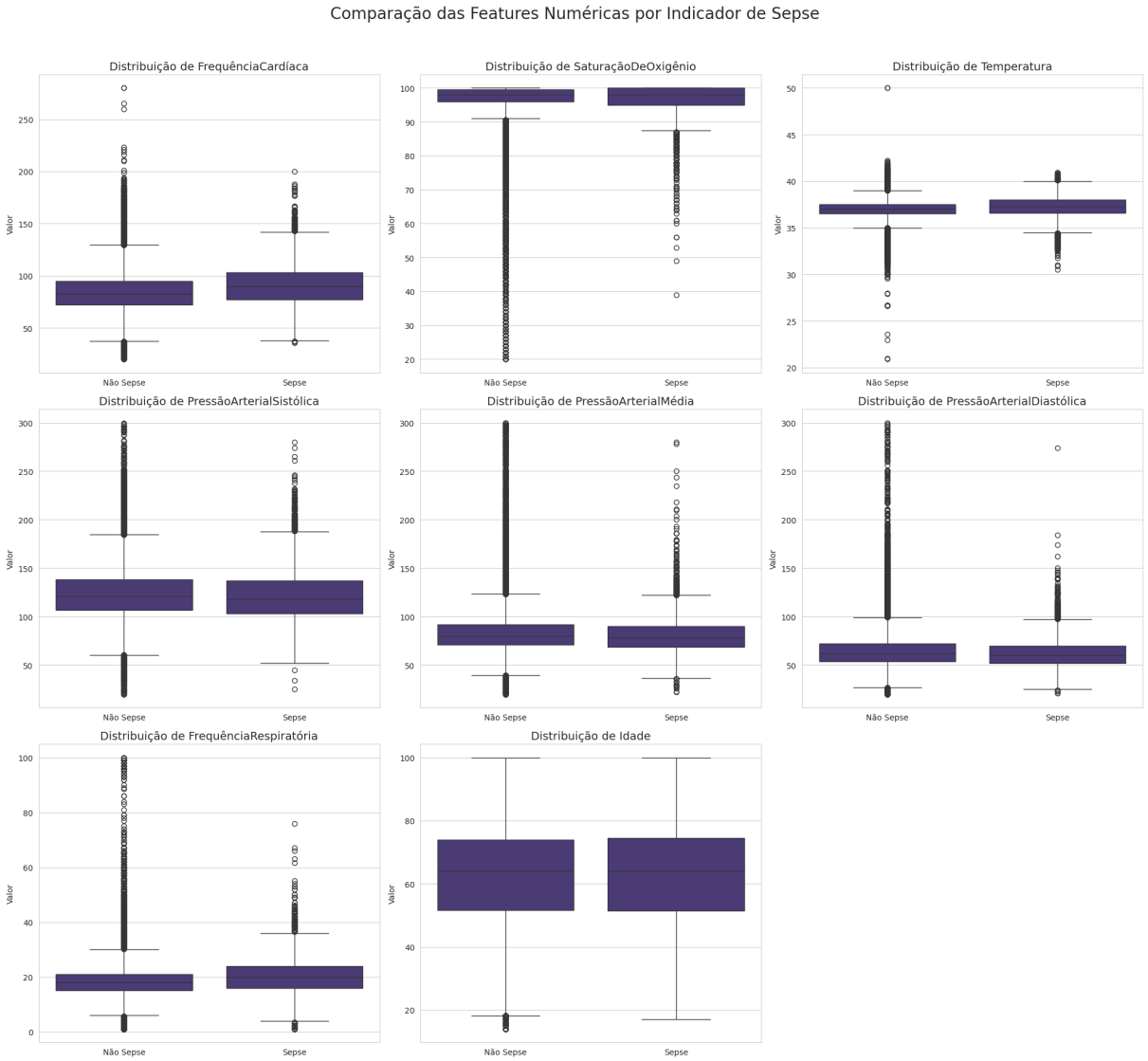

Figura 6 – Boxplots comparativos entre pacientes positivos e negativos

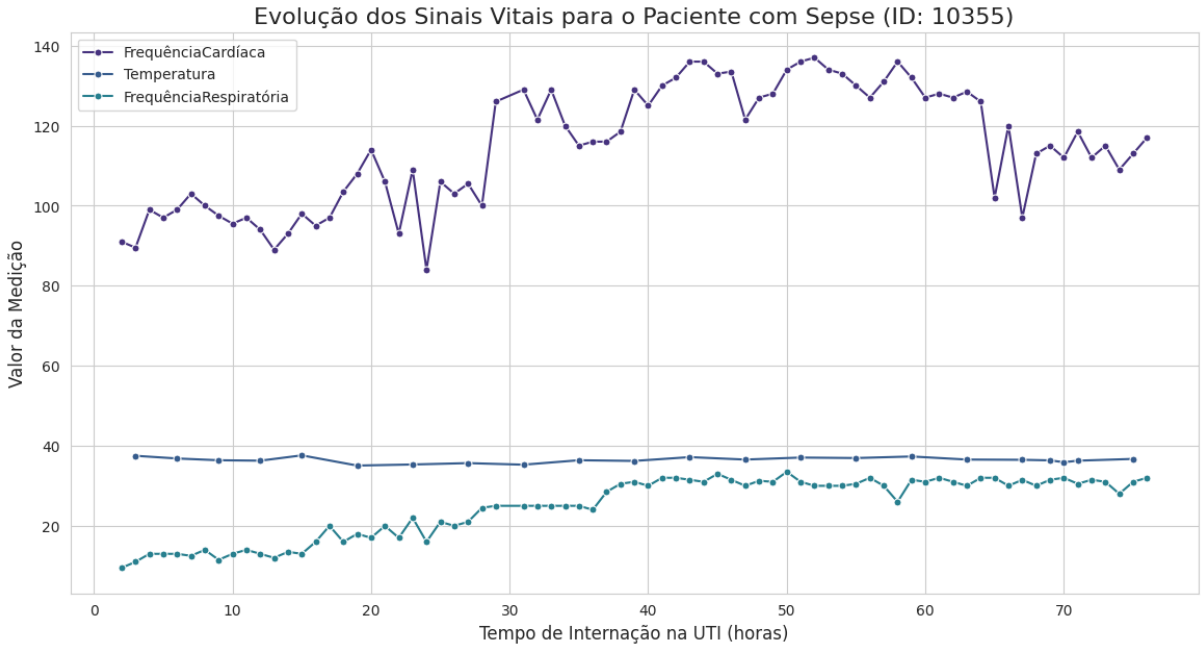

Figura 7 – Gráfico de linha para um paciente positivo aleatório

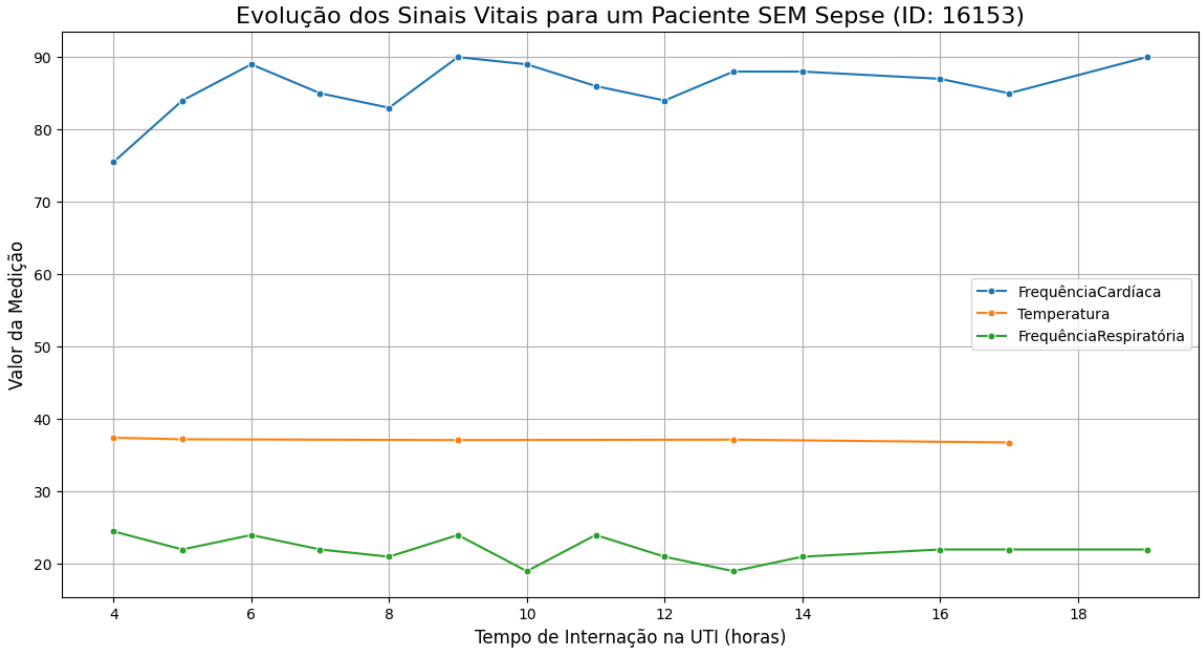

Figura 8 – Gráfico de linha para um paciente negativo aleatório

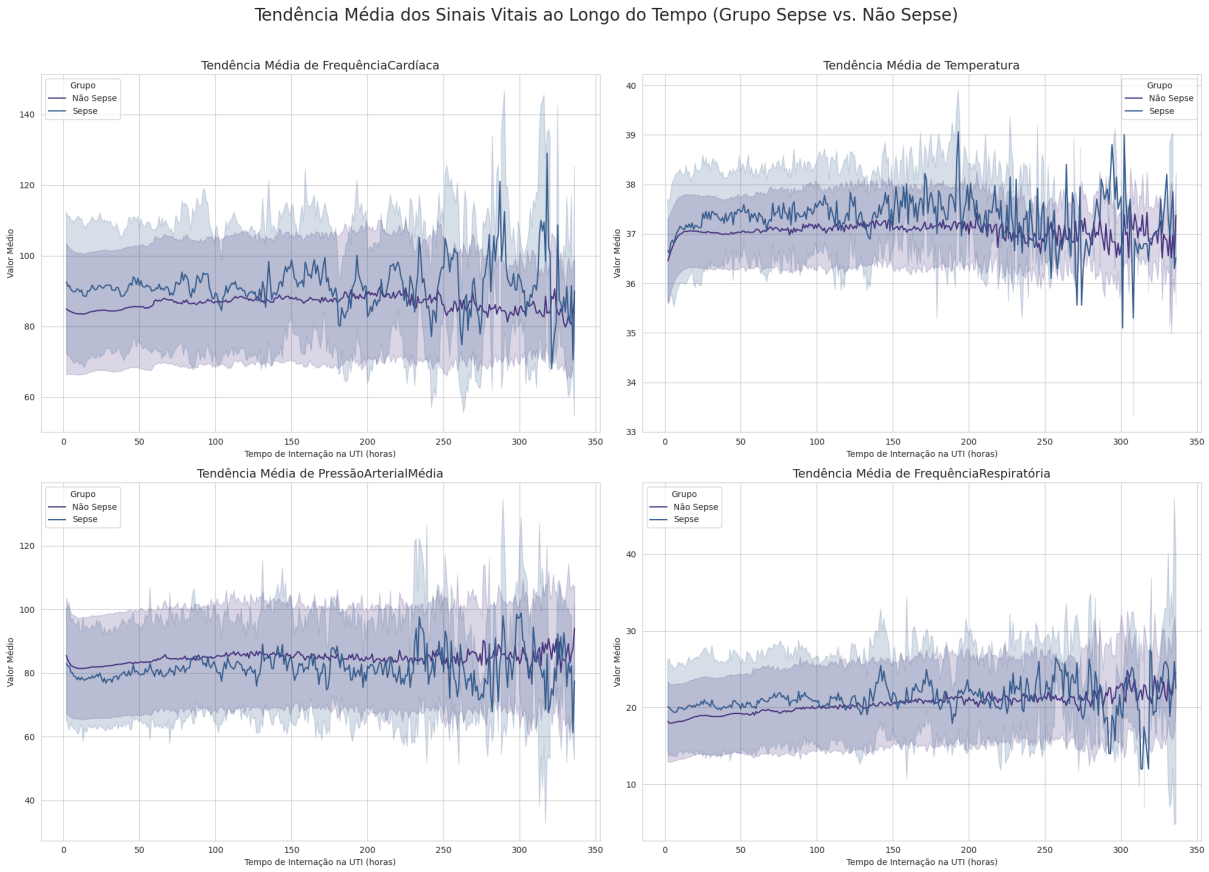

Figura 9 – Série temporal de atributos de pacientes com e sem sepse (comparativa)

Facilmente, ainda na Figura 9, nota-se que há padrões distintos e tal não estão focados apenas no tempo final na UTI, há tendências padronizadas distintas ainda no inicio, aparentemente relacionadas a uma inclinação maior a taquicardia, temperaturas que elevam-se mais rapidamente e pressão baixa, além disso é gritante conforme o avanço do tempo mobilizado na UTI as quedas abruptas na frequência respiratória. Tais indicadores sendo mapeados são coerentes com os sinais da doença. Tais tendências também não são sútis o bastante para não serem identificadas pelos gráficos Q-Q com uma certa facilidade, esses revelam linearmente a contribuição em comparativo com uma régua normal, conforme é possível verificar nas Figura 10 e 11. A distribuição da população de cada coluna expõe preenchimentos que reafirmam as séries temporais da Figura 9, por exemplo os registros da frequência cardiaca em pacientes com o quadro povoam mais facilmente valores mais baixos assim como a frequência respiratória.

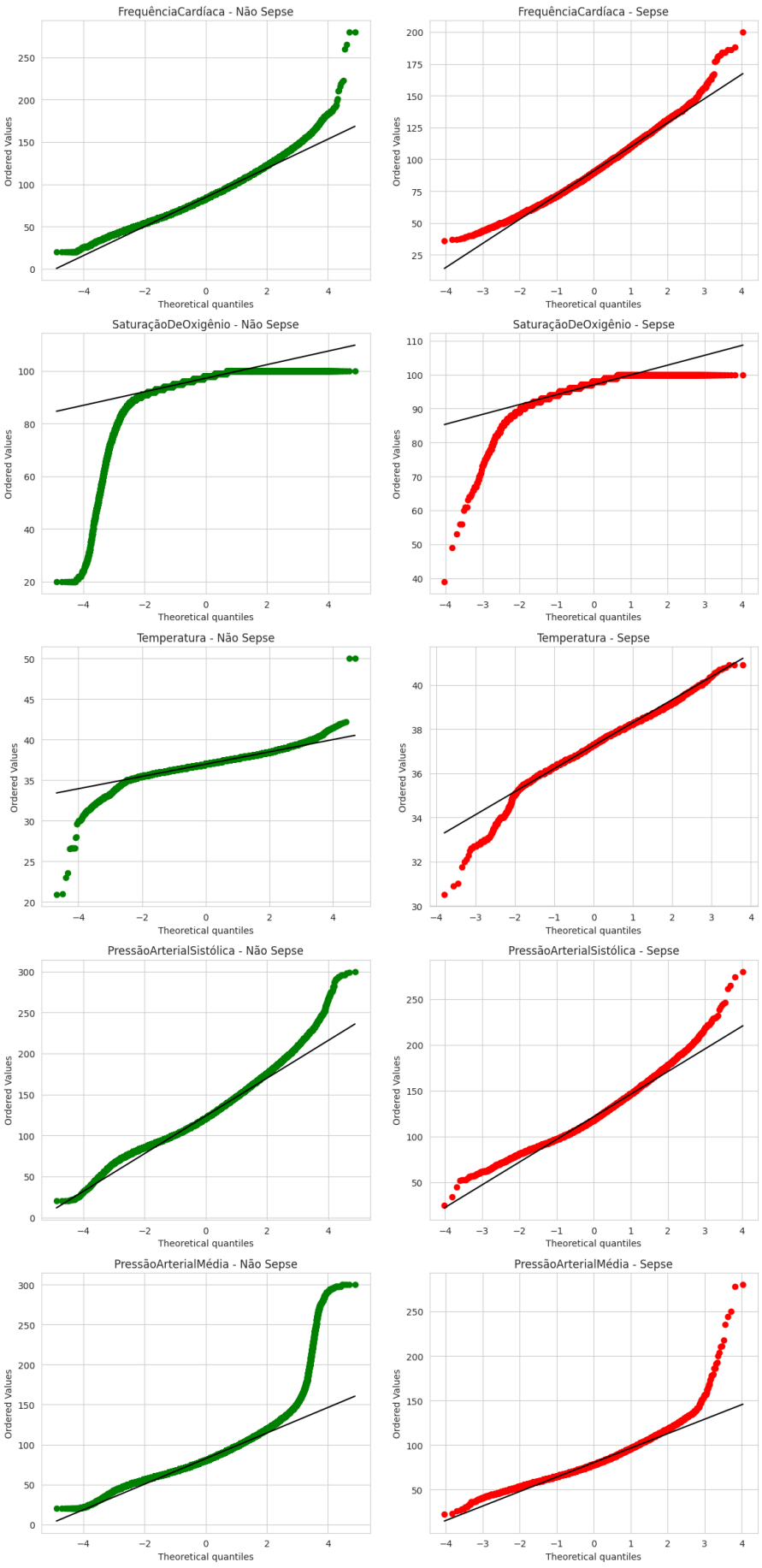

Figura 10 – Distribuição das colunas em gráficos Q-Q

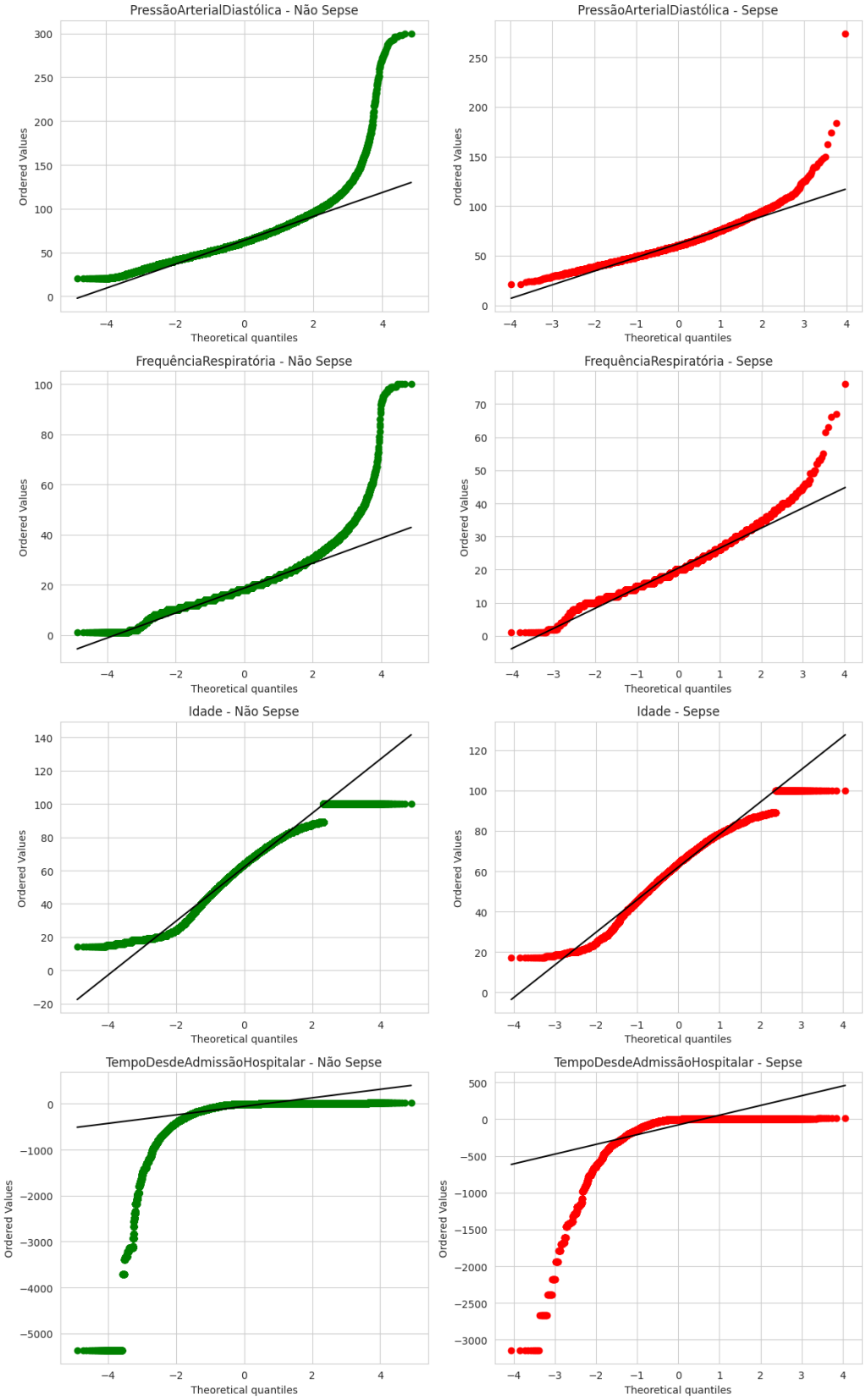

Figura 11 – Distribuição das colunas em gráficos Q-Q

1.2.5 Potenciais Ajustes Necessários

Há uma problemática relacionada à predição, devido ao fato de que diversas linhas da base de dados referem-se a um mesmo paciente. Além disso, a variável alvo pode mudar ao longo do tempo conforme ocorrem eventos específicos no histórico do paciente. Por exemplo, é possível que um determinado paciente tenha as primeiras seis entradas com a variável alvo igual a 0, e a partir da sétima entrada essa variável torne-se 1. Para tratar essa questão e tornar a predição mais eficaz, há uma abordagem em perspectiva:

- Agrupar todas as entradas de um mesmo paciente numa única linha, utilizando estatísticas como máximos, mínimos e médias das colunas numéricas ao longo do tempo;
- Incluir informações sobre ocorrências de valores anormais, como batimentos cardíacos elevados ou baixa oxigenação no sangue, como indicadores binários;
- Caso em alguma das entradas agrupadas para um paciente a variável alvo (sepse) seja igual a 1, o valor final da variável alvo para essa linha agregada também será definido como 1.

Ainda é válido ressaltar que outros esforços visiveis tratam-se de lidar com a nulidade das variáveis e consolidar o conjunto de dados utilizados para treino e para testes. Outro fator significativa refere-se ao balanceamento de classes diante da diferença gritante entre a quantia apresentada e a normalização dos valores para uma formatação manipulável e legível para os modelos.

1.3 Preparação dos Dados

Antecipadamente é válido ressaltar que muitos processos de processamento envolvidos foram resolvidos ainda na etapa de exploração, isso devido ao fato da necessidade alarmante de transformações para que houvesse sentido. Além disso, tal ato é coeso com o CRISP que trata-se de uma metodologia não fixa ao processo linear.

1.3.1 Seleção dos Dados

A Tabela 2 evidencia um cenário crítico em relação à completude dos dados: apenas 15 das 43 colunas possuem taxa de preenchimento superior a 60%, enquanto a maioria apresenta valores nulos em mais de 90% dos registros. Embora o critério convencional para a seleção de atributos em análises robustas costume adotar um limite mínimo de 70% de preenchimento, tal exigência resultaria em uma base consideravelmente empobrecida, limitada a apenas 11 variáveis. Assim, optou-se por um limiar mais

flexível de 60%, permitindo a retenção de 15 atributos considerados potencialmente informativos. Além disso, dada a relevância clínica da variável *temperatura*, mesmo com apenas 33,84% de preenchimento, sua inclusão foi justificada mediante a perspectiva de uso de estratégias de imputação — como propagação do último valor observado (forward fill) ou preenchimento com constantes clínicas típicas. Dessa forma, o conjunto de dados a ser trabalhado foi definido com 16 colunas. Adicionalmente, foi realizado o processo de renomeação das colunas para o idioma português com termos clínicos mais compreensíveis, o que favorece a leitura e interpretação dos resultados. A Tabela 2 já apresenta os nomes atualizados das variáveis conforme essa padronização. É válido ressaltar que após as formatações necessárias para melhor ingestão dos algoritmos, as quais reverberaram numa base mais diminuta e adequada (processo que será explicado mais detalhadamente na 1.3.5) foi selecionado sobre a base final 80% do seus registros para procedimento de treino e 20% para testagem, a fim de avaliar o desempenho com dados inéditos, dessa forma trazendo percepção a possiveis overfittings (desempenho eficaz apenas diante do aprendizado dos ruidos de um modelo, ou seja, "decorar"). Houve estratificação, assim mantendo uma proporção de classes igualitária entre os conjuntos.

1.3.2 Limpeza dos Dados

Foi gerada uma versão filtrada da base contendo apenas os pacientes com diagnóstico de sepse. Contudo, a Tabela 3 mostra que o padrão de valores ausentes permaneceu praticamente inalterado (e, em alguns casos, com aumento) o que indica que a ocorrência da sepse não está associada a maior completude de nenhuma variável específica. Em relação à duplicidade, observou-se comportamento esperado no contexto de séries temporais: é comum a repetição de valores em colunas clínicas entre medições consecutivas. No entanto, não foram encontradas duplicações de linhas inteiras, tampouco conflitos entre combinações de ID e hora, reforçando a integridade dos dados. Por fim, campos como idade e sexo mantêm-se consistentes dentro de cada grupo de identificação (ID), mesmo com múltiplos registros ao longo do tempo, o que reforça a coerência estrutural da base após a reestruturação. Foi realizada uma análise da estrutura e formatação dos dados com o intuito de identificar possíveis inconsistências ou necessidades de correção antes da fase de modelagem. Observou-se que a base apresenta um nível satisfatório de qualidade estrutural. Nenhuma coluna apresenta tipagens mistas (valores de diferentes tipos coexistindo na mesma variável), o que é um indicativo positivo de coerência no armazenamento dos dados. Além disso, variáveis categóricas de natureza binária, como *Sexo* e *IndicadorDeSepse*, demonstram valores distintos compatíveis com sua proposta, apresentando exatamente duas categorias válidas e sem anomalias aparentes, como valores ausentes, inconsistentes ou fora do domínio esperado. No que se refere às unidades de medida, foi possível constatar

que as colunas seguem uma uniformidade aparente. A presença dessas unidades de forma padronizada colabora para evitar erros durante transformações posteriores, como normalizações ou comparações entre variáveis. As variáveis mantêm o mesmo significado ao longo dos registros, sem alterações de nomenclatura, ambiguidade ou mudanças contextuais nos dados. Não foram detectados valores evidentemente impossíveis para os dados clínicos (por exemplo, temperaturas negativas ou frequências cardíacas incompatíveis com a vida humana). A positiva carência de anomalias pode ser visualizada na Tabela 6. Também não há indícios de que registros de um mesmo paciente apresentem conflitos entre informações fixas, como idade ou sexo, garantindo que a estrutura de série temporal por indivíduo seja preservada com fidelidade.

1.3.3 Construção dos Dados

A partir da estatística sumária evidenciada na Tabela 5 e Tabela 6 foi adotada uma abordagem cuidadosa e multifásica para o tratamento de valores ausentes, de modo a preservar a integridade das informações clínicas e garantir a coerência estatística necessária para a modelagem. Dado que os registros representam séries temporais por paciente, optou-se por uma estratégia de imputação sequencial baseada no histórico individual de cada paciente.

Inicialmente, os valores ausentes foram preenchidos utilizando o método de *forward fill* e *backward fill*, ou seja, preenchendo os campos vazios com o último valor conhecido ou com o valor subsequente, desde que pertencentes ao mesmo paciente (identificado pela coluna `ID`). Essa abordagem busca manter a coesão dos dados clínicos ao longo do tempo, respeitando a individualidade de cada trajetória hospitalar.

Nos casos em que não havia dados prévios ou posteriores disponíveis para realizar tal preenchimento, foi aplicada uma segunda etapa de imputação, dessa vez com base em estatísticas globais da base de dados. Para variáveis quantitativas contínuas, como sinais vitais e exames laboratoriais, utilizou-se a média aritmética da respectiva coluna como substituição. Já para variáveis categóricas ou discretas, empregou-se a moda, ou seja, o valor mais frequente, a fim de manter a coerência com os padrões predominantes do conjunto.

Após o tratamento dos valores ausentes, realizou-se uma transformação estrutural dos dados para viabilizar a modelagem supervisionada. Como a base original se constitui de registros por hora de acompanhamento dos pacientes, a granularidade temporal foi agregada por paciente. Para isso, os dados foram agrupados com base no identificador único de paciente (`ID`), e as variáveis contínuas foram transformadas em estatísticas descritivas agregadas, especificamente os valores de **mínimo**, **máximo** e **média** ao longo do tempo de internação.

Essa agregação tem como objetivo sintetizar a trajetória clínica de cada indivíduo

em uma única instância de entrada para o modelo de aprendizado de máquina, preservando os aspectos dinâmicos dos sinais vitais e exames por meio de estatísticas representativas. Além disso, essa abordagem reduz a dimensionalidade da base e facilita a interpretação dos resultados preditivos posteriormente obtidos.

1.3.4 Integração dos Dados

Durante a etapa de preparação dos dados, identificou-se uma limitação crítica: a ausência da coluna de identificação dos pacientes na base originalmente disponibilizada, o que comprometia a segmentação correta das informações individuais ao longo da série temporal. Essa limitação dificultava o agrupamento adequado dos registros por paciente, sendo que cada linha representa um momento distinto (em horas) de monitoramento clínico. Para contornar essa deficiência estrutural, foi realizada uma busca por versões alternativas do mesmo conjunto de dados e encontrou-se uma versão hospedada no Kaggle (Soni, 2022), que mantinha a coluna "ID"original. Após validações cuidadosas de consistência entre os datasets, foi possível realizar a reintegração dessa variável ao conjunto de dados principal, permitindo mapear cada sequência de horas a um paciente específico de forma precisa. Essa ação não apenas aumentou a confiabilidade da análise como também otimizou substancialmente o tempo de processamento, evitando a necessidade de técnicas heurísticas para deduzir artificialmente os agrupamentos com base em reinícios da contagem da coluna "Hour". Portanto, a reinserção da variável de identificação foi um passo decisivo para garantir a qualidade das análises subsequentes, respeitando a estrutura sequencial dos dados clínicos e possibilitando uma aplicação correta de técnicas baseadas em séries temporais.

1.3.5 Formatação dos Dados

Além do tratamento padrão de dados relatado acima, a preparação estendeu-se visando o propósito de aprendizagem dos modelos. As três ações protagonistas nesse cenário foram o balanceamento das duas classes target, dado que inicialmente a classe 0 que refere-se a pessoas não diagnosticadas com quadro sepse ocupava 94.4% em contrapartida complementar a classe 1, pacientes com sepse, apenas 5.6%. As estratégias utilizadas foram SMOTE (OverSampling da classe minoritária, elevando-a para ocupação de 70%) para criar exemplos sintéticos da classe minoritária (classe 1, sepse), aumentando sua ocupação e UnderSampling da classe majoritária (classe 0) reduzindo o número de pacientes sem sepse baseado-se em criterização aleatória para tal diminuição, mantendo uma apenas 70%. Além disso, outra ação de relevância, para fins de otimização de treinamento e coerência com o modelo, foi a compactação dos registros, tendo em vista que um paciente representava vários eventos/linhas, assim trazendo as medidas estatisticas nas colunas necessárias resumiu-se numa numeração equivalente a cardinalidade dos pacientes via paciente ID gerando 21.055

linhas. Por fim, sobre essa base final de dados foi a normalização desses registros alterando suas tipagens enquanto mantém-se seu significado. Nessa normalização nos dados númericos (a maioria das colunas) foi utilizado o standard scaler, transformando os registros para que tenham média 0 e desvio padrão 1. Enquanto nas categóricas abstraiu-se com one-hot encoder, transformando-as em colunas numéricas codificadas equivalentemente e evitando ordem inexistentes, assim no processo de ML trabalha-se apenas com campos númericos. Além disso, houve aplicações para evitar data leakage, aplicando as transformações (como padronização e codificação) dentro da pipeline, ou seja, somente após separar os dados em treino/teste, dessa maneira o modelo não aprende o conjunto de teste (que deve ser totalmente "desconhecido") trazendo um aprendizado enganosamente bom.

1.4 Modelagem

A fase de modelagem consiste na seleção, aplicação e ajuste de algoritmos supervisionados para a criação de modelos preditivos. Para isso, são considerados critérios como: capacidade de generalização, robustez a dados desbalanceados, interpretabilidade e complexidade computacional. A seguir, são descritas as técnicas selecionadas para experimentação. A respeito do treinamento, do modelo de busca, da avaliação e de preparações prévias relevantes é possível visualizar os códigos no notebook público (Allyson, 2025). Acompanha-se no tópico o detalhamento à respeito da otimização dos modelos com o fine-tuning de todos os modelos listados, incluindo o uso de técnicas de busca por hiperparâmetros em todos os modelos com validação cruzada como especificado, resultados em métricas como recall, F-1 score, matriz de confusão e outras, inclusive em gráficos, aprofundamento do desempenho em treinamento, teste (dado testes de estresse com dados novos e desconhecidos para avaliar a robustez dos modelos) e validação, uma perspectiva análitica de custo-benefício e detalhes sobre a complexidade de tais como número de parâmetros e tempo de inferência.

1.4.1 Seleção da Técnica de Modelagem

- **K-NN (K-Nearest Neighbors)**: Classifica um novo ponto com base na maioria dos seus vizinhos mais próximos. Eficiente em datasets pequenos e bem balanceados, mas sensível a escalas e ao desbalanceamento de classes. A expectativa é que depende do quão qualificado foi o balanceamento de classes;
- **LVQ (Learning Vector Quantization)**: Variante supervisionada do k-means, que cria protótipos representativos por classe. Funciona com dados normalizados e é interpretável, mas menos robusto a ruídos;

- **Árvore de Decisão**: Cria regras hierárquicas com base nos atributos mais informativos. Boa interpretabilidade, mas suscetível a overfitting em datasets ruidosos. Combinada com podas ou em florestas (como Random Forest), torna-se poderoso;
- **SVM (Support Vector Machine)**: Tenta encontrar um hiperplano ótimo que separa as classes. É eficaz em espaços de alta dimensionalidade, especialmente com uso de *kernels*. O uso é promissor no caso, desde que balanceamento e tuning cuidadoso tenham a qualidade esperada;
- **Random Forest**: Conjunto de árvores de decisão treinadas com amostragem aleatória dos dados. Robusto, lida bem com dados originalmente desbalanceados e fornece estimativas de importância dos atributos. É um dos modelos esperados com melhor performance para o cenário;
- **Rede Neural MLP (Multilayer Perceptron)**: Modelo neural com múltiplas camadas densamente conectadas. Pode capturar relações complexas e não-lineares, mas requer balanceamento e tuning muito cuidadoso de hiperparâmetros;
- **Comitê de Redes Neurais Artificiais**: Vários MLPs com decisões combinadas. Reduz variância e aumenta robustez, mas apresenta elevado custo computacional. Adequado para cenários com muitos dados e modelos instáveis;
- **Comitê Heterogêneo (Stacking)**: Combinação de diferentes algoritmos (ex: SVM, árvore, rede neural) cujas predições são usadas como entrada para um modelo final. Muito poderoso, especialmente quando os modelos base são diversos e bem selecionado, necessita ser idealmente orquestrado;
- **XGBoost**: Modelo de *boosting* de árvores altamente eficiente. Lida bem com desbalanceamento, possui regulação embutida e é muito utilizado em competições. Forte candidato para melhor resultado.
- **LightGBM**: Variante do XGBoost com maior eficiência computacional. Escala bem para grandes conjuntos de dados e também lida bem com desbalanceamento. Recomendado especialmente se o tempo de treinamento for uma preocupação. Assim como o XGBoost, as expectativas são altas.

1.4.2 Geração do Modelo

A geração do modelo corresponde à aplicação prática dos algoritmos de aprendizado supervisionado selecionados previamente, com o objetivo de construir preditores iniciais baseados nos dados de treino. Esta etapa visa obter um primeiro conjunto de modelos treinados, ainda sem ajustes finos de parâmetros (hiperparâmetros), de modo a permitir comparações iniciais de comportamento e viabilidade.

É importante ressaltar que a fase de geração é distinta da calibração: aqui o foco está no treinamento com a configuração padrão ou com busca aleatória inicial, enquanto a otimização fina de desempenho será abordada na próximo tópico. Para garantir consistência, boas práticas como a separação treino-teste com estratificação e o uso de *pipelines* foram empregadas como já citado anteriormente.

1.4.2.1 Estrutura de Treinamento

Todos os modelos foram treinados com os seguintes cuidados metodológicos já relatados na subseção 1.3.5:

- **Divisão dos dados**: 80% para treino e 20% para teste, com **estratificação da variável alvo** (SepsisLabel) para manter a proporção original das classes, que estão fortemente desbalanceadas.
- **Pipeline de pré-processamento**: Incluiu padronização com *StandardScaler* para variáveis numéricas e codificação *OneHotEncoder* para variáveis categóricas. Essas transformações foram aplicadas exclusivamente sobre o conjunto de treino e replicadas no teste, evitando *data leakage*.
- **Validação cruzada**: Para garantir robustez, a avaliação preliminar dos modelos usou *StratifiedKFold* com 5 dobras (`n_splits=5`), respeitando a distribuição original das classes.
- **Treinamento paralelo e aleatório**: Os modelos foram gerados com auxílio do `RandomizedSearchCV`, com 20 iterações por algoritmo e paralelização total (`n_jobs=-1`) para maior eficiência;
- **Reprodutibilidade**: Todos os experimentos foram executados com seeds fixas (`random_state=42`), garantindo reprodutibilidade dos resultados.

1.4.2.2 Resultados iniciais

Foram gerados os seguintes modelos preditivos, considerando suas características individuais:

- **Random Forest**: Treinado com 20 configurações aleatórias iniciais, demonstrou elevada robustez à presença de ruídos e desbalanceamento, sendo especialmente promissor para este problema clínico;
- **Árvore de Decisão**: Utilizada como baseline simples e interpretável. Por ser suscetível a sobreajuste, sua eficácia isolada tende a ser limitada;

- **K-Nearest Neighbors (KNN)**: Embora útil em contextos com dados bem distribuídos, mostrou desempenho inferior nos testes preliminares devido à sensibilidade ao desbalanceamento;
- **Nearest Centroid (como alternativa similar ao LVQ)**: Classificador simples baseado em distância ao centroide de cada classe. Usado pela ausência de implementação padrão do LVQ na biblioteca scikit-learn;
- **Support Vector Machine (SVM)**: Inicialmente treinada com kernels lineares e RBF. Embora robusta em espaços de alta dimensão, sua sensibilidade ao desbalanceamento impõe cuidados adicionais, dando ainda mais ênfase na fase de balanceamento anterior.
- **Multilayer Perceptron (MLP)**: Rede neural de múltiplas camadas, avaliada com 18 configurações de arquitetura distintas (como número de camadas, neurônios, função de ativação e regularização), totalizando 90 ajustes via validação cruzada. Apresenta potencial para capturar padrões complexos nos dados;
- **Comitê de MLPs**: Ensemble de múltiplos MLPs combinados para gerar uma predição mais robusta. Foram testadas 12 variações de comitês, com 60 ajustes no total. Destaca-se pela capacidade de reduzir variância e instabilidade de redes isoladas;
- **XGBoost**: Modelo baseado em árvores de decisão sequenciais com boosting, conhecido por seu alto desempenho em competições. Foram avaliadas 20 combinações de hiperparâmetros (profundidade, número de árvores, learning rate), totalizando 100 ajustes. Apresenta excelente desempenho, mesmo com dados desbalanceados;
- **LightGBM**: Variante otimizada do gradient boosting, mais leve e veloz, também avaliada com 20 configurações (100 ajustes no total). Indicado para grandes volumes de dados e conjuntos com muitas features;
- **Comitê Heterogêneo (Stacking)**: Técnica de ensemble que combina previsões de diferentes tipos de modelos (por exemplo, SVM, RF, MLP) em uma etapa final de metaclassificação. Neste estudo, o stacking foi utilizado com configuração fixa, sem ajuste de hiperparâmetros, mas buscando tirar proveito da complementaridade entre modelos fortes.

1.4.2.3 Observações Iniciais

Nesta etapa, o modelo **Random Forest** já apresentou métricas promissoras, como alto *AUC-ROC* e bom *F1-score* na validação cruzada, mesmo sem otimizações avançadas. Modelos como **KNN** e **LVQ** tiveram desempenho mais limitado, indicando maior sensibilidade ao desbalanceamento e à escala dos dados.

Com a introdução de modelos mais robustos, observou-se melhora significativa nos resultados. A **Rede Neural MLP** apresentou boa capacidade de generalização, especialmente ao ser combinada em comitê (ensemble), o que reduziu variações entre os folds. Entretanto, os melhores desempenhos foram alcançados pelos modelos de **gradient boosting**, especialmente **XGBoost** e **LightGBM**, que superaram os demais em métricas como *AUC-ROC* e *F1-score*. Esses modelos mostraram-se mais eficazes em capturar padrões complexos e lidar com o desbalanceamento, mantendo desempenho consistente e eficiente. Apesar disso, enfatiza-se que são mais onerosos que o Random Forest.

O modelo **Stacking**, mesmo sem busca de hiperparâmetros, também obteve resultados competitivos ao combinar previsões de modelos distintos, reforçando o potencial dos comitês heterogêneos neste tipo de tarefa. Essas observações destacam a superioridade dos métodos baseados em boosting neste cenário clínico, justificando sua priorização nas etapas seguintes de análise e comparação, contudo tendo em perspectiva a qualidade do Random Forest mediante a custo computacional.

1.4.3 Calibração dos Parâmetros

Visando maximizar o desempenho preditivo dos modelos com ajuste fino busca-se os hiperparâmetros e reavaliação em múltiplas métricas. Essa fase é essencial para melhorar a capacidade preditiva, uma vez que parâmetros padrões raramente representam o melhor ajuste para dados reais. Antecipadamente, deve-se visualizar quais são os hiperparâmetros, conforme nas Tabelas 7, 8 e 9.

Tabela 7 – Modelos e Espaços de Busca de Hiperparâmetros

Modelo	**Espaço de Busca**
RandomForest	n_estimators: [50–300], max_depth: [3–20], min_samples_split: [2–10]
DecisionTree	max_depth: [3–20], min_samples_split: [2–10]
KNN	n_neighbors: [3–15], weights: [’uniform’, ’distance’]
LVQ (Centroid)	Não possui hiperparâmetros
SVM	C: distribuição contínua, kernel: [’rbf’, ’linear’]

Tabela 8 – Modelos Mais Complexos e Espaços de Busca de Hiperparâmetros

Modelo	**Espaço de Busca**
MLP	hidden_layer_sizes: [(50,), (100,), (50, 50)], activation: ['relu', 'tanh'], alpha: [1e-5, 1e-4, 1e-3],
Comitê MLP	n_estimators: [5, 10, 20], hidden_layer_sizes: [(50,), (100,)], alpha: [1e-4, 1e-3]
XGBoost	n_estimators: [50, 100, 200], max_depth: [3, 5, 7], learning_rate: [0.01, 0.1, 0.3]
LightGBM	n_estimators: [50, 100, 200], max_depth: [3, 5, 7], learning_rate: [0.01, 0.1, 0.3]

Tabela 9 – Configuração Padrão do Modelo Stacking

Elemento	**Descrição**
Modelo Base 1	MLPClassifier ('hidden_layer_sizes': 50, max_iter: 300, random_state:42)
Modelo Base 2	XGBClassifier('use_label_encoder': False, 'eval_metric': 'logloss', 'random_state': 42)
Modelo Base 3	LGBMClassifier('random_state':42)
Meta-Modelo Final	LogisticRegression()
Parâmetro passthrough	True (atributos originais também são passados ao meta-modelo)

1.4.3.1 Método de Busca

Como dito no tópico 1.4.2.1 o método de busca para o ajuste dos hiperparâmetros foi adotado o **RandomizedSearchCV**, que consiste em amostrar aleatoriamente combinações dentro de um espaço de busca predefinido. Essa abordagem foi escolhida por permitir uma exploração eficiente com menor custo computacional, comparada ao Grid Search.

- **Iterações por modelo:** 20 combinações diferentes por algoritmo.
- **Validação cruzada:** *StratifiedKFold* com 5 dobras, respeitando a distribuição das classes durante a busca.
- **Execução paralela:** Utilização de `n_jobs=-1` para acelerar o processo por meio de paralelização total.

1.4.3.2 Espaço de Busca

Os espaços de busca foram definidos com base em ranges realistas para cada modelo, conforme mostrado na Tabela 7 anteriormente.

1.4.4 Melhores Hiperparâmetros Encontrados

Após a execução do processo de calibração, foram identificadas as melhores combinações de hiperparâmetros para cada modelo, conforme apresenta a Tabela 10.

Tabela 10 – Melhores Hiperparâmetros por Modelo

Modelo	**Melhores Hiperparâmetros**
RandomForest	{'model__max_depth': 16, 'model__min_samples_split': 3, 'model__n_estimators': 299}
DecisionTree	{'model__max_depth': 6, 'model__min_samples_split': 9}
KNN	{'model__n_neighbors': 14, 'model__weights': 'distance'}
LVQ (Centroid)	(sem hiperparâmetros)
SVM	{'model__C': 0.9847, 'model__kernel': 'rbf'}
MLP	{'model__hidden_layer_sizes': 50, 'model__alpha': 0.001, 'model__activation': 'relu'}
Comitê MLP	{'model__n_estimators': 20, 'model__hidden_layer_sizes': 100, 'model__alpha': 0.0001}
XGBoost	{'model__n_estimators': 200, 'model__max_depth': 7, 'model__learning_rate': 0.1}
LightGBM	{'model__n_estimators': 200, 'model__max_depth': 7, 'model__learning_rate': 0.1}
Stacking	Configurações default

1.4.5 Resultados Pós-Calibração

A calibração evidenciou ganhos de desempenho relevantes para todos os modelos, com destaque para o Random Forest, que combinou robustez e estabilidade nas métricas obtidas em múltiplas iterações. Modelos como KNN e SVM também apresentaram melhoria com os ajustes, especialmente na sensibilidade, fundamental para a detecção precoce da sepse.

Adicionalmente, os modelos baseados em gradiente, como XGBoost e LightGBM, demonstraram os melhores desempenhos globais, tanto em AUC-ROC quanto em F1-score, sendo especialmente eficazes no tratamento do desbalanceamento após o pré-processamento com SMOTE e subamostragem. O MLP também apresentou resultados satisfatórios após otimização da arquitetura, e o Comitê MLP (via Bagging) contribuiu para maior estabilidade e redução de variância nas previsões.

O modelo Stacking, mesmo sem busca sistemática por hiperparâmetros, se destacou ao combinar a força de diferentes algoritmos base (MLP, XGBoost, LightGBM) com um meta-modelo de regressão logística, apresentando desempenho competitivo e equilibrado entre sensibilidade e especificidade. Assim, os experimentos evidenciam o benefício da diversidade de abordagens no aumento da capacidade preditiva, especialmente em cenários clínicos críticos como o da sepse.

1.5 Avaliação dos Modelos

Após a calibração, os modelos foram avaliados num conjunto de teste estratificado e previamente separado (20% dos dados originais), não visto durante o treinamento. As métricas consideradas foram: Acurácia, Precisão, Recall, F1-score e AUC-ROC, com especial atenção à sensibilidade (recall), dado o contexto crítico de detecção de sepse.

1.5.1 Métricas de Avaliação

A Tabela 11 apresenta o desempenho dos modelos no conjunto de teste.

Tabela 11 – Desempenho dos Modelos no Conjunto de Teste

Modelo	Accuracy	Precision	Recall	F1-score	AUC-ROC	Matriz de Confusão
RandomForest	0.9570	0.8509	0.4957	0.6265	0.8946	[7138, 342] [206, 381]
DecisionTree	0.9531	0.8119	0.4634	0.5900	0.8389	[6858, 622] [208, 379]
KNN	0.9405	0.8452	0.2232	0.3531	0.7849	[6365, 1115] [245, 342]
LVQ	0.7351	0.1327	0.4770	0.2076	0.6230	[5538, 1942] [300, 287]
SVM	0.9435	0.8291	0.2811	0.4198	0.8628	
MLP	0.889812	0.927789	0.889812	0.468619	0.851532	[6787, 695] [194, 392]
Cômite MLP	0.908775	0.929386	0.908775	0.5	0.872339	[6964, 518] [218, 368]
XGBoost	0.947323	0.947867	0.947323	0.64135	0.90009	[6964, 518] [218, 368]
LightGBM	0.950545	0.950429	0.950545	0.658683	0.907949	[7284, 198] [201, 385
Stacking	0.941373	0.946492	0.941373	0.629601	0.896973	[7193, 289] [184, 402]

1.5.2 Overfitting e Underfitting

Analisando os resultados obtidos, observa-se que modelos como Random Forest, XGBoost e LightGBM apresentam métricas de teste elevadas (AUC-ROC acima de 0,89 e F1-score acima de 0,62), sugerindo boa capacidade de generalização e baixo risco de overfitting. Entretanto, o LVQ apresenta desempenho significativamente inferior (AUC-ROC de 0,623 e F1-score de 0,2076), o que indica provável underfitting, possivelmente decorrente de limitações intrínsecas do algoritmo para dados de alta dimensionalidade e complexidade.

Modelos como KNN e SVM, embora possuam AUC-ROC relativamente altos, sofrem com recall muito baixo, sugerindo dificuldade em identificar corretamente casos positivos, o que pode estar relacionado mais a desequilíbrios na sensibilidade do classificador do que a overfitting propriamente dito. Já redes neurais como MLP e Comitê MLP apresentam desempenho moderado, sem indícios claros de sobreajuste, mas ainda aquém dos modelos de gradiente.

Dessa forma, não há evidências de overfitting severo nos melhores modelos, mas há sinais claros de underfitting em abordagens mais simples ou inadequadas para a natureza dos dados, como LVQ e, em menor escala, Decision Tree.

1.5.3 Análise Comparativa

A partir dos resultados, observa-se que:

- **Random Forest** apresentou o melhor desempenho geral. Seu AUC-ROC (0,8946) e F1-score (0,6265) foram os mais altos entre os modelos base, combinando robustez à heterogeneidade dos dados com um bom equilíbrio entre sensibilidade (Recall = 0,4957) e precisão (Precision = 0,8509);
- **Decision Tree** foi competitivo, com bom desempenho em todas as métricas e alta interpretabilidade. Apresentou um AUC-ROC de 0,8389 e F1-score de 0,59. Pode ser indicado em contextos onde a transparência do modelo é prioritária;
- **SVM** obteve um AUC-ROC de 0,8628, o que é bom, mas seu recall baixo (0,2811) e F1-score de apenas 0,4198 sugerem que, apesar de acurácia geral elevada, não consegue detectar adequadamente os casos de sepse.
- **KNN** teve boa precisão (0,8452), mas um recall extremamente baixo (0,2232), o que compromete sua utilidade em detecção precoce, onde identificar corretamente os casos positivos (sepse) é crucial;
- **LVQ** apresentou o pior desempenho global, com AUC-ROC de 0,623, evidenciando underfitting severo. O recall de 0,477 indica alguma detecção de positivos, mas com baixíssima precisão (0,1327), o que resulta em muitos falsos positivos;
- **MLP** obteve um F1-score de 0,4686 e AUC-ROC de 0,8515, com precisão de 0,9278 e recall de 0,8898, além de acurácia de 0,8898. Embora apresente boa capacidade de identificar corretamente os casos positivos (recall alto), o valor relativamente baixo do F1-score sugere que o modelo ainda enfrenta desafios para equilibrar essa detecção com a precisão no contexto clínico, podendo gerar falsos positivos que demandariam investigação médica adicional. Matriz de Confusão: [6787, 695] [194, 392];
- **Comitê MLP** alcançou um F1-score de 0,5000 e AUC-ROC de 0,8723, com precisão de 0,9294, recall de 0,9088 e acurácia de 0,9088. O uso do ensemble reduziu variações de desempenho observadas no MLP individual, aumentando tanto a estabilidade quanto a robustez. Isso é particularmente relevante para dados clínicos heterogêneos, onde o ganho em recall é benéfico para a detecção precoce;
- **XGBoost** apresentou um F1-score de 0,6414 e AUC-ROC de 0,9001, com alta precisão (0,9479), recall (0,9473) e acurácia (0,9473). Esse desempenho consistente em todas as métricas demonstra forte capacidade de generalização e

modelagem de padrões complexos, além de bom equilíbrio entre sensibilidade e especificidade, o que é ideal para contextos onde tanto a detecção quanto a redução de falsos alarmes são essenciais. Matriz de Confusão: [6964, 518] [218, 368];

- **LightGBM** foi o modelo com o melhor desempenho geral, com F1-score de 0,6587, AUC-ROC de 0,9079, precisão de 0,9504, recall de 0,9505 e acurácia de 0,9505. Sua alta performance aliada à velocidade de execução o torna particularmente adequado para aplicações clínicas em tempo real, onde decisões rápidas e precisas podem impactar significativamente o prognóstico do paciente. Matriz de Confusão: [7284, 198] [201, 385];
- **Stacking** obteve F1-score de 0,6296 e AUC-ROC de 0,8970, com precisão de 0,9465, recall de 0,9414 e acurácia de 0,9414. Ao combinar diferentes modelos base, o stacking conseguiu superar a maioria dos algoritmos individuais, apresentando equilíbrio entre as métricas e se beneficiando da diversidade dos classificadores, o que o torna uma opção robusta para cenários de diagnóstico precoce. Matriz de Confusão: [7193, 289] [184, 402];

1.5.3.1 Estatística Descritiva dos Parâmetros

A fim de visualização mais ágil do resultado exponhe-se alguns gráficos seletos abaixo. A Figura 12 apresenta uma comparação direta das métricas Accuracy, Precision, Recall, F1-score e AUC-ROC para cada modelo avaliado. Observa-se que LightGBM, XGBoost e Random Forest mantêm desempenho elevado e consistente em todas as métricas, reforçando sua robustez e capacidade de generalização. O modelo LVQ apresenta desempenho significativamente inferior, sobretudo em F1-score, Recall e AUC-ROC, evidenciando dificuldade em capturar corretamente os padrões da base. O KNN mantém boa Precision, mas sofre acentuadamente no Recall, comprometendo sua utilidade para detecção de casos positivos. O Comitê MLP apresenta um equilíbrio mais favorável entre Precision e Recall em relação à MLP isolada, indicando ganhos provenientes do uso de ensemble. Já o Random Forest, além da consistência nas métricas, combina alta Precision e AUC-ROC com Recall competitivo, sendo uma alternativa sólida ao lado dos modelos de gradiente.

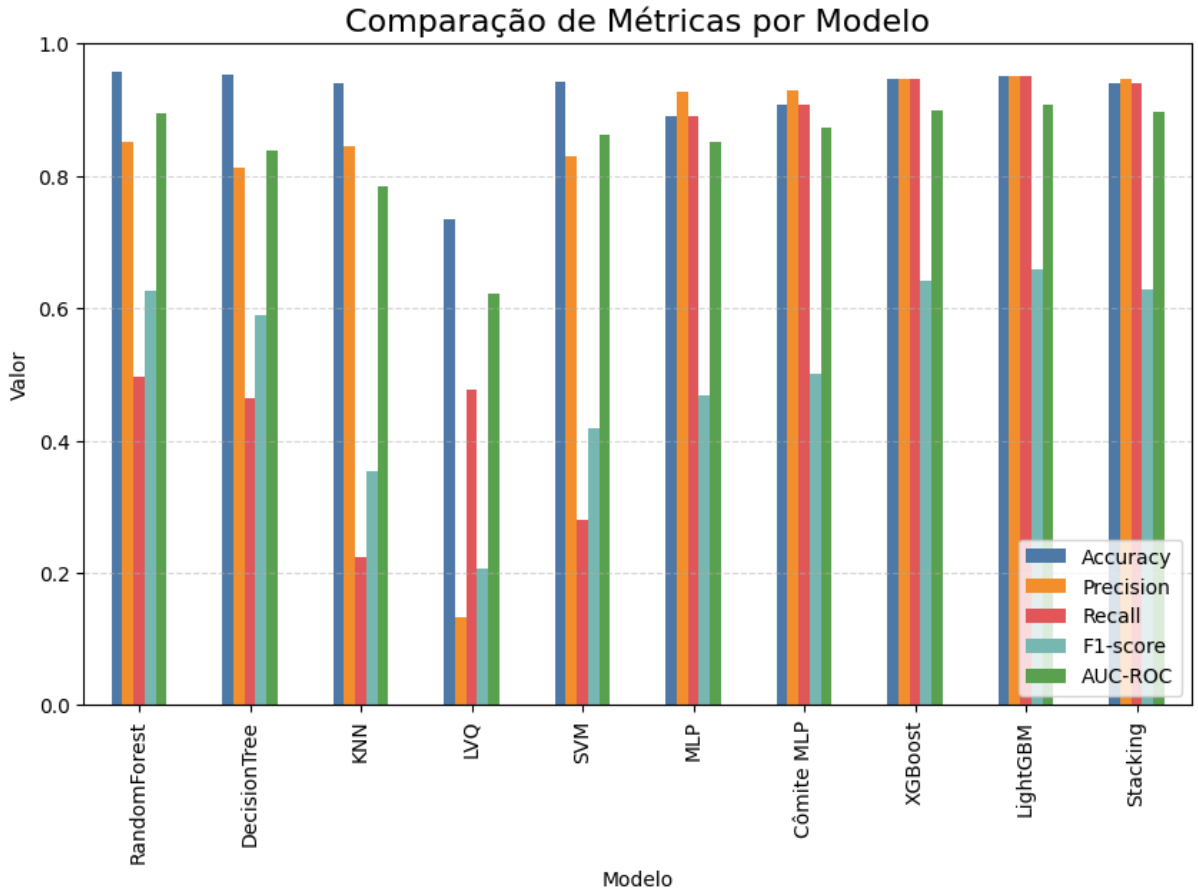

Figura 12 – Comparação direta das métricas de avaliação dos modelos treinados.

A Figura 13 detalha a evolução das métricas durante a otimização, oferece uma visão sobre a consistência de cada modelo. Observa-se que LightGBM e XGBoost seguem uma trajetória de melhora estável a cada nova configuração, indicando uma otimização eficaz. Em contraste, o Random Forest e o Stacking se destacam por atingir um platô de alta performance de forma precoce e com notável consistência, reforçando sua robustez. A análise evolutiva também expõe as fragilidades de modelos como o LVQ, que se mostra anômalo e instável com métricas erráticas, e do KNN e SVM, que falham em evoluir na métrica crítica de Recall, comprometendo sua aplicabilidade prática.

Na Figura 14 o formato permite observar o equilíbrio de desempenho entre as métricas entre os modelos mais simples. O Random Forest apresenta um perfil mais abrangente (preenchendo melhor o gráfico), enquanto modelos como LVQ e KNN mostram regiões com baixa performance, especialmente em recall.

Figura 13 – Evolução das métricas durante treinamento com as 20 melhores configurações.

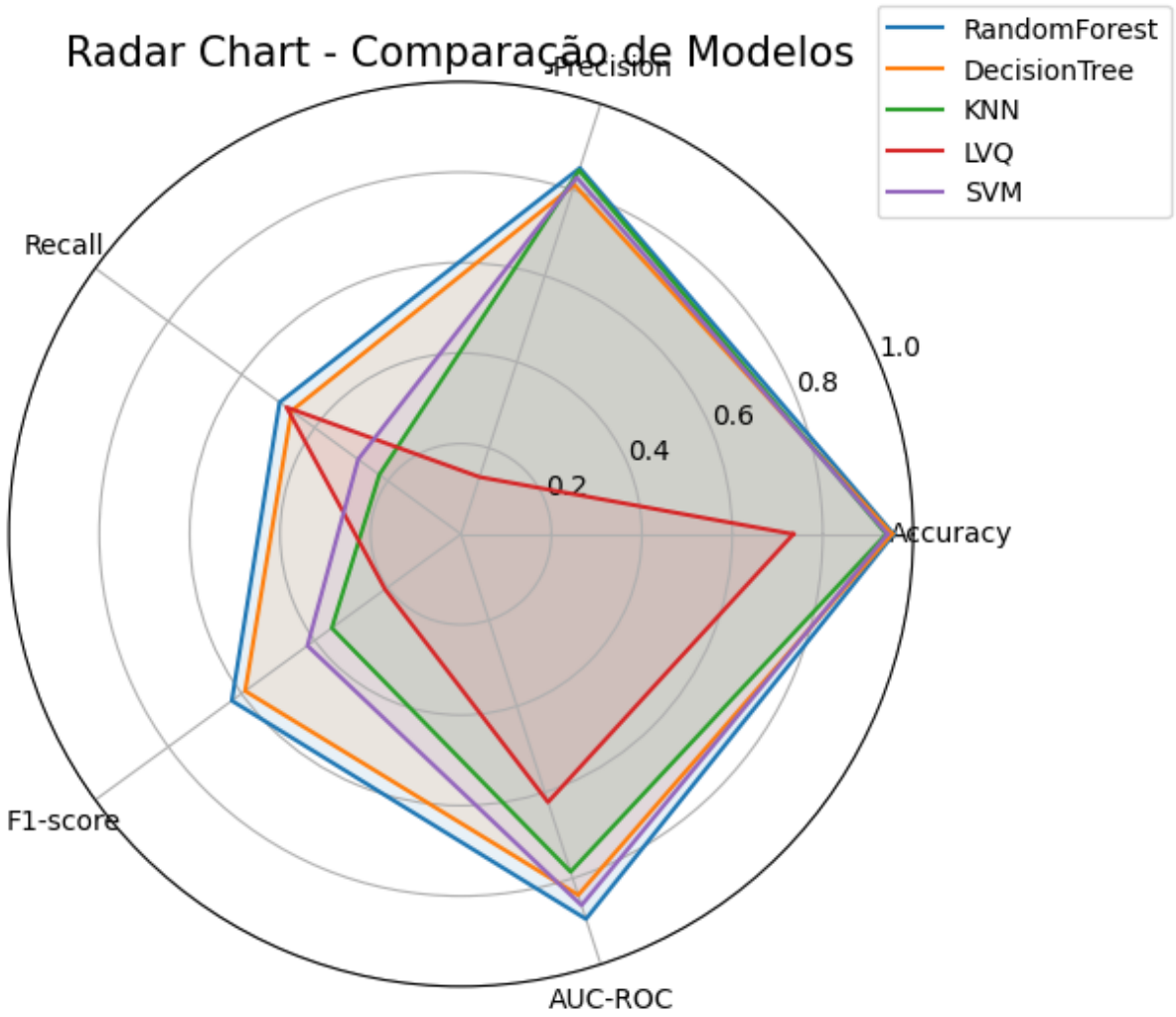

Figura 14 – Visualização em radar das métricas dos modelos.

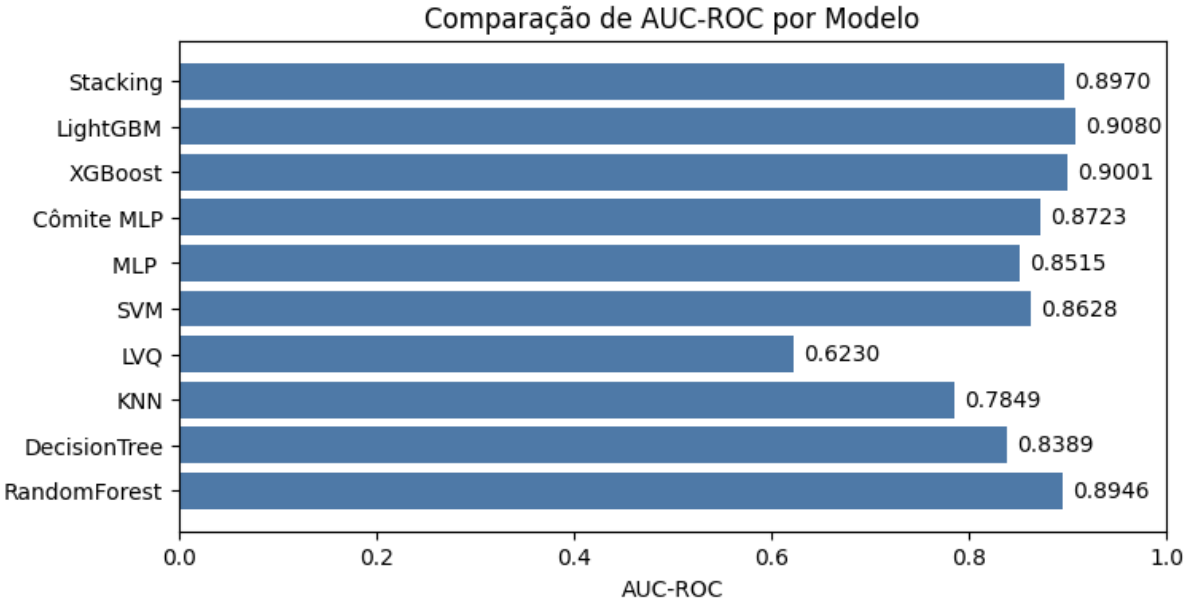

Figura 15 – AUC-ROC por modelo.

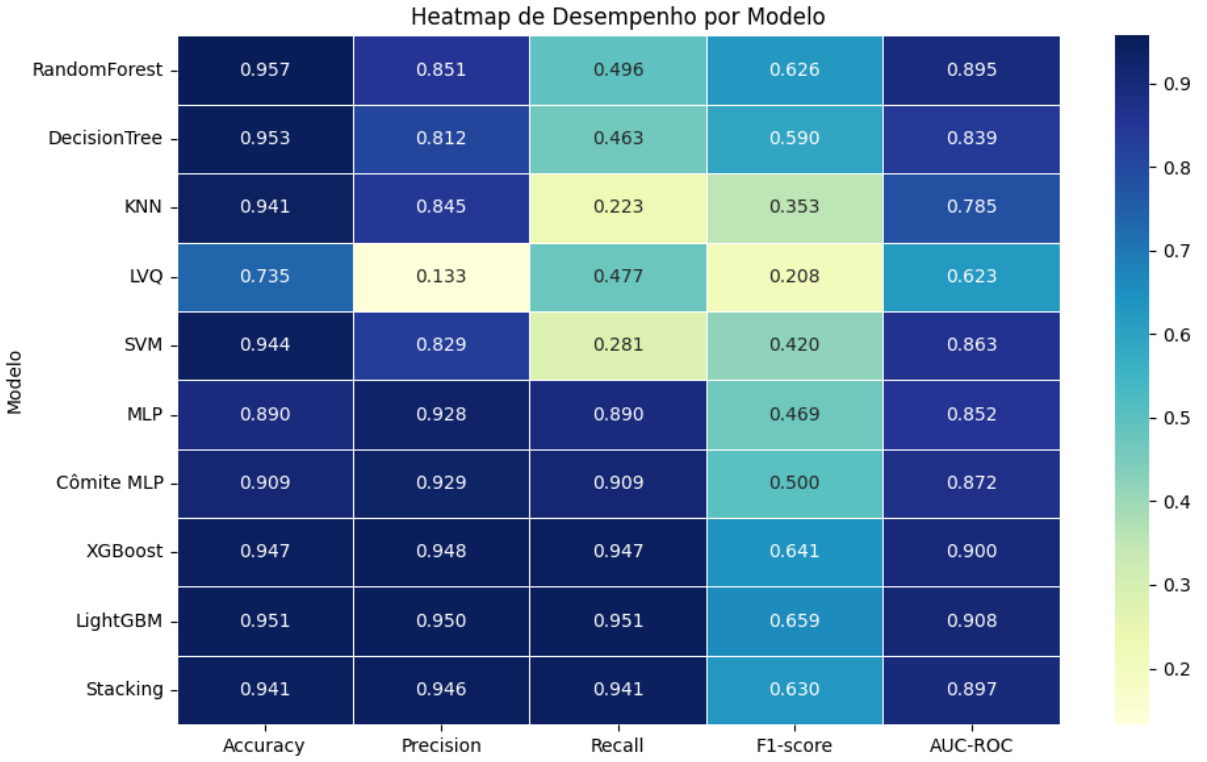

Figura 16 – fig: Heatmap comparando todos os Modelos.

Novamente na Figura 15, observa-se que os modelos LightGBM, XGBoost e Stacking lideram com valores de AUC-ROC superiores a 0.89, indicando excelente capacidade de separação entre as classes. O Comitê MLP e a SVM também mantêm desempenhos competitivos. Por outro lado, o modelo LVQ permanece significativamente abaixo dos demais (0.623), evidenciando sua limitação em discriminar corretamente as classes.

Por fim, a Figura 16 apresenta uma visão condensada do desempenho dos modelos por meio de um heatmap, evidenciando a superioridade dos ensembles (Random Forest, XGBoost, LightGBM e Stacking) em múltiplas métricas. O LVQ apresenta os piores resultados, especialmente em Precision e F1-score, enquanto KNN e SVM sofrem com Recall reduzido. MLP e Comitê MLP têm Precision e Recall altos, mas menor equilíbrio no F1-score. O Stacking mantém desempenho competitivo e estável, mesmo sem extensa otimização, explorando a complementaridade de diferentes algoritmos. O LightGBM supera o Stacking em Recall e F1-score, sendo mais indicado quando a prioridade é detectar o máximo de casos positivos. Já o Stacking destaca-se pela robustez e boa adaptação a diferentes cenários.

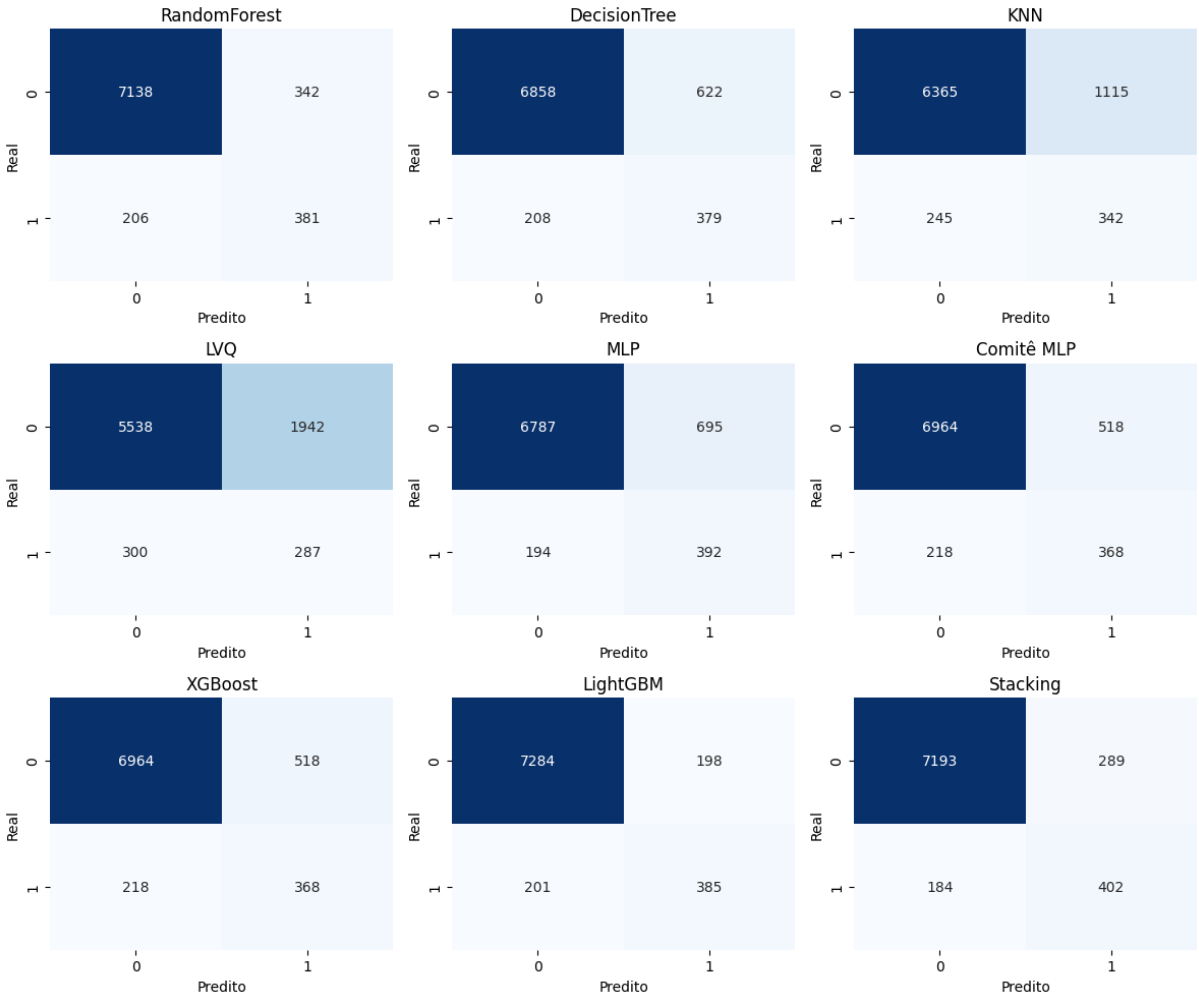

Figura 17 – fig: Matriz de confusão de cada modelo.

A Figura 17 apresenta as matrizes de confusão dos modelos, permitindo avaliar a distribuição de acertos e erros entre as classes. Nota-se que LightGBM apresenta o menor número de falsos negativos (201), o que reforça seu alto Recall e sua capacidade de identificar casos positivos de sepse, característica crucial para diagnóstico precoce. O Stacking também demonstra bom equilíbrio, com poucos falsos negativos (184) e falsos positivos relativamente controlados, reforçando sua robustez geral. Já o Random Forest e o XGBoost mantêm desempenho sólido, embora apresentem ligeiro aumento de falsos positivos em comparação ao LightGBM. Em contrapartida, modelos como KNN e, principalmente, LVQ sofrem com taxas elevadas de falsos positivos, o que reduz sua precisão e compromete a confiabilidade das previsões. O MLP e o Comitê MLP exibem distribuição mais equilibrada de erros, mas ainda ficam aquém dos melhores ensembles na redução simultânea de falsos positivos e negativos. Esses resultados confirmam que, no contexto de detecção de sepse, minimizar falsos negativos é prioritário, favorecendo modelos como LightGBM, Stacking e XGBoost, apesar da interpretabilidade e resultado considerável do Random Forest.

1.5.3.2 Considerações Finais

A avaliação evidenciou a importância do balanceamento prévio e do uso de validação cruzada com estratificação. Além disso, destaca-se:

- A importância da métrica *recall*, associada à sensibilidade clínica, modelos com baixo recall podem não detectar pacientes com sepse, o que é indesejável;
- A capacidade do **Random Forest** de aliar desempenho competitivo a alta interpretabilidade, fator crucial na governança de dados sensíveis, permitindo explicabilidade das decisões e suporte a auditorias internas e externas;
- A boa performance da MLP e, principalmente, do **Comitê MLP**, que demonstraram precisão elevada e sensibilidade consistente, mostrando-se promissores na detecção de padrões complexos de dados clínicos;
- A eficácia de modelos ensemble como XGBoost, LightGBM e Stacking, que também apresentaram métricas robustas, embora com diferentes graus de complexidade e requisitos computacionais;
- A identificação, por meio de busca sistemática, de configurações mais eficazes para os modelos testados, especialmente no caso do Random Forest e dos modelos baseados em gradiente.

1.6 Implementação

A fase de implementação corresponde ao momento em que o modelo desenvolvido deixa o ambiente experimental e passa a ser integrado a um contexto real, no qual suas previsões podem de fato auxiliar no diagnóstico precoce da sepse. Entre as opções avaliadas, o **Random Forest** desponta como a alternativa mais indicada para produção, não apenas por seu desempenho sólido, mas principalmente por oferecer vantagens únicas no campo da *governança de dados sensíveis*. Sua natureza baseada em múltiplas árvores de decisão permite uma *explicabilidade* clara sobre como as previsões são obtidas, facilitando a interpretação clínica, a rastreabilidade de decisões e a conformidade com regulamentações e políticas de auditoria em saúde.

O Random Forest também apresenta baixo risco de sobreajuste quando configurado adequadamente, mantém estabilidade frente a ruídos nos dados e possui custo computacional acessível, permitindo execução rápida mesmo em infraestruturas hospitalares sem alto poder de processamento. Essa combinação de robustez técnica, interpretabilidade e viabilidade operacional o torna a escolha preferencial para ambientes clínicos que demandam confiança, transparência e rastreabilidade.

Ainda que o Random Forest seja a recomendação principal, outros modelos se destacam em contextos específicos. O **Comitê MLP**, por exemplo, mostrou-se eficaz na captura de relações não lineares complexas, beneficiando-se da combinação de múltiplas redes neurais para melhorar a generalização. A **MLP** individual, embora menos robusta que o comitê, apresentou métricas consistentes e pode ser vantajosa quando se busca simplicidade de arquitetura com boa capacidade de modelagem de padrões temporais e não lineares. O **LightGBM** e o **XGBoost** também permanecem como opções viáveis, especialmente em cenários que demandam maximização de *recall* e adaptação a bases de dados maiores e mais heterogêneas.

1.6.1 Planejamento da Implementação e Integração

Para a implantação do Random Forest, é imprescindível replicar integralmente o pipeline de pré-processamento utilizado no treinamento, preservando a consistência dos dados de entrada. A integração com sistemas de prontuário eletrônico deve ser feita por meio de interfaces seguras, capazes de enviar automaticamente as variáveis necessárias e retornar probabilidades de sepse de forma clara, por exemplo, via alertas visuais ou notificações diretas à equipe médica. Além disso, a interpretabilidade nativa do Random Forest permitirá a geração de justificativas legíveis para cada previsão, reforçando a confiança da equipe clínica no sistema.

1.6.2 Monitoramento e Manutenção

O monitoramento contínuo deve priorizar a manutenção do *recall* e o controle da taxa de falsos positivos, equilibrando sensibilidade e especificidade. A capacidade do Random Forest de fornecer importâncias de variáveis será utilizada para detectar mudanças no perfil dos dados (*data drift*), permitindo ajustes proativos. Ciclos periódicos de reavaliação e re-treinamento serão essenciais para garantir a atualização do modelo frente a mudanças epidemiológicas ou de protocolos hospitalares.

1.6.3 Geração de Relatórios

A geração de relatórios técnicos e clínicos será facilitada pela estrutura interpretável do Random Forest, permitindo documentar não apenas métricas de desempenho e versões do modelo, mas também explicar, em termos clínicos, as razões que levaram a cada alerta emitido. Isso potencializa a capacidade de auditoria, fortalece a governança de dados e aumenta a aceitação do sistema por parte das equipes médicas.

1.7 Resumo Executivo

O presente trabalho teve como objetivo desenvolver e avaliar modelos de aprendizado de máquina para detecção precoce de sepse em pacientes hospitalizados, utilizando

a metodologia CRISP-DM como guia. A análise indicou o **Random Forest** como o modelo mais indicado para implementação prática, por combinar desempenho competitivo, baixo custo computacional e, sobretudo, interpretabilidade e explicabilidade — características essenciais na gestão de dados sensíveis em saúde.

A **MLP** e o **Comitê MLP** também se destacaram, especialmente pela boa capacidade de modelar padrões complexos, enquanto modelos baseados em gradiente (LightGBM e XGBoost) ofereceram excelente sensibilidade e precisão. No entanto, a escolha pelo Random Forest se justifica pela facilidade de auditoria, rastreabilidade das decisões e robustez frente a variações nos dados, tornando-o a solução mais segura e transparente para apoio ao diagnóstico clínico.

1.8 Limitações

As limitações permanecem relacionadas à qualidade dos dados, desbalanceamento entre classes, risco de perda de generalização para outros contextos hospitalares e restrições computacionais em modelos mais complexos. Mesmo o Random Forest, apesar de interpretável, pode perder desempenho se houver mudanças bruscas no perfil dos dados ou se o pré-processamento não for fielmente replicado. Além disso, a ausência de acompanhamento clínico contínuo durante todas as fases do projeto pode ter limitado a priorização de variáveis mais relevantes.

1.9 Trabalhos Futuros

Para evoluir, propõe-se a ampliação da base de dados para múltiplos centros, implementação de técnicas de explicabilidade complementares como SHAP e LIME, e integração de um sistema de *learning* contínuo com governança rigorosa. Pretende-se também avaliar a aplicação do Random Forest e do Comitê MLP em outras condições críticas, explorando o potencial desses modelos em um ecossistema mais amplo de apoio à decisão clínica baseado em dados.

REFERÊNCIAS

ALLYSON. *Resultados de Modelos de Machine Learning - PH4.1*. 2025. <https://colab.research.google.com/drive/1mRZ26NkemGK_Hvy49vHI6KrnmXbzoIIL?usp=sharing>. Acessado em: 10 jul. 2025. Citado na página 30.

ARAúJO, S.; SILVA, C. Sistema de apoio à decisão clínica com uso de machine learning e ciência de dados na identificação da sepse. *Anais do Encontro Nacional de Inteligência Artificial e Computacional (ENIAC)*, Sociedade Brasileira de Computação, v. 2023, p. 285–296, 2023. Disponível em: <https://sol.sbc.org.br/index.php/eniac/article/view/33795>. Citado na página 7.

Centers for Disease Control and Prevention (CDC). *Sepsis - About CDC's Sepsis Program*. 2024. <https://www.cdc.gov/sepsis/about/index.html>. Acesso em: 20 maio 2025. Citado na página 6.

COSTA, B. N. d. *Predição de sepse em unidades de terapia intensiva utilizando aprendizado de máquina*. 2022. <https://repositorio.ufcspa.edu.br/items/56dba9fd-bd78-455e-b7f1-270180849944>. Trabalho de Conclusão de Curso (Graduação) — Universidade Federal de Ciências da Saúde de Porto Alegre (UFCSPA). Citado na página 7.

Drauzio Varella. *Sepse (septicemia)*. 2025. <https://drauziovarella.uol.com.br/doencas-e-sintomas/sepse-septicemia>. Acesso em: 20 maio 2025. Citado na página 6.

KORE, S. *Exploratory Data Analysis (EDA)*. 2023. <https://medium.com/@snehalkore016/exploratory-data-analysis-eda-d30967cb5589>. Acesso em: 11 jun. 2025. Citado na página 10.

MOOR, M.; RIECK, B.; HORN, M.; JUTZELER, C. R.; BORGWARDT, K. *Early Prediction of Sepsis in the ICU Using Machine Learning: A Systematic Review*. 2019. Disponível em: <https://arxiv.org/abs/1906.02956>. Citado na página 7.

PABBA, K. *A Comprehensive Guide on Exploratory Data Analysis (EDA)*. 2023. <https://medium.com/@pabbakavya123/a-comprehensive-guide-on-exploratory-data-analysis-eda-ab38f33d6abc>. Acesso em: 11 jun. 2025. Citado na página 10.

PhysioNet. *PhysioNet/Computing in Cardiology Challenge 2019: Early Prediction of Sepsis from Clinical Data*. 2019. <https://physionet.org/content/challenge-2019/1.0.0/>. Acesso em: 20 maio 2025. Citado na página 10.

SONI, L. *Sepsis Prediction*. 2022. Kaggle Notebook. Disponível em: <https://www.kaggle.com/code/lakshyasoni97/sepsis-prediction>. Citado 3 vezes nas páginas 7, 10 e 29.