



Dissertação para Mestrado

Proposta e Avaliação de um Modelo de Prognóstico para Pacientes com Septicemia

Marcos Vinícius Silva Bento
mvsb@ic.ufal.br

Orientadores:

Dr. Rafael de Amorim Silva
Dr. Bruno Pimentel

Maceió
Agosto 26, 2022

Marcos Vinícius Silva Bento

Proposta e Avaliação de um Modelo de Prognóstico para Pacientes com Septicemia

Dissertação apresentada ao curso de Mestrado em Informática do Programa de Pós Graduação em Informática da Universidade Federal de Alagoas, como requisito para obtenção do grau de Mestre em informática.

Orientadores:

Dr. Rafael de Amorim Silva

Dr. Bruno Pimentel

Maceió
Agosto 26, 2022

Catálogo na Fonte
Universidade Federal de Alagoas
Biblioteca Central
Divisão de Tratamento Técnico

Bibliotecário: Marcelino de Carvalho Freitas Neto – CRB-4 - 1767

B478p Bento, Marcos Vinícius Silva.
Proposta e avaliação de um modelo de prognóstico para pacientes com septicemia / Marcos Vinícius Silva Bento. – 2022.
72 f. : il.

Orientador: Rafael de Amorim Silva.
Co-orientador: Bruno Pimentel.
Dissertação (mestrado em informática) - Universidade Federal de Alagoas. Instituto de Computação. Maceió, 2022.

Bibliografia: f. 62-67.
Apêndices: f. 68-72.

1. Prognóstico. 2. Diagnóstico. 3. Sepsis. 4. Unidades de terapia intensiva. 5. Aprendizagem de máquina. I. Título.

CDU: 004.81:159.953.5:616.94



UNIVERSIDADE FEDERAL DE ALAGOAS/UFAL
Programa de Pós-Graduação em Informática – PPGI
Instituto de Computação/UFAL
Campus A. C. Simões BR 104-Norte Km 14 BL 12 Tabuleiro do Martins
Maceió/AL - Brasil CEP: 57.072-970 | Telefone: (082) 3214-1401



Folha de Aprovação

MARCOS VINÍCIUS SILVA BENTO

PROPOSTA E AVALIAÇÃO DE UM MODELO DE PROGNÓSTICO PARA PACIENTES COM SEPTICEMIA

Dissertação submetida ao corpo docente do Programa de Pós-Graduação em Informática da Universidade Federal de Alagoas e aprovada em 26 de agosto de 2022.

Banca Examinadora:

Documento assinado digitalmente
 RAFAEL DE AMORIM SILVA
Data: 26/10/2022 09:45:25-0300
Verifique em <https://verificador.iti.br>

Prof. Dr. RAFAEL DE AMORIM SILVA
UFAL – Instituto de Computação
Orientador

Documento assinado digitalmente
 BRUNO ALMEIDA PIMENTEL
Data: 30/08/2022 10:04:23-0300
Verifique em <https://verificador.iti.br>

Prof. Dr. BRUNO ALMEIDA PIMENTEL
UFAL – Instituto de Computação
Coorientador

Documento assinado digitalmente
 DIEGO DERMEVAL MEDEIROS DA CUNHA M/
Data: 29/08/2022 14:26:47-0300
Verifique em <https://verificador.iti.br>

Prof. Dr. DIEGO DERMEVAL MEDEIROS DA CUNHA MATOS
UFAL – Instituto de Computação

Examinador Interno
Documento assinado digitalmente

 ALMIR PEREIRA GUIMARAES
Data: 29/08/2022 21:29:36-0300
Verifique em <https://verificador.iti.br>

Prof. Dr. ALMIR PEREIRA GUIMARÃES
UFAL – Instituto de Computação
Examinador Externo

Com todo meu amor. Por tudo que fez e faz por mim, a minha amada Mãe, Sicleide dos Santos Silva.

Dedico.

Agradecimentos

A minha amada Mãe Sicleide, que fez de sua prioridade a minha educação, a melhor herança que poderia contruir ao longo dos anos.

A todos os familiares que de alguma forma torceram durante essa caminhada, em especial ao meu primo Lucas, por ser uma fonte de inspiração na busca por conhecimento, e me apoiou desde o início.

Aos meus amigos, que comemoraram e torceram por mim nessa caminhada, em especial *Duda*, que dividiu algumas experiências no decorrer do mestrado.

Aos colegas do Centro de Pesquisa em Tecnologia Emergente, Maxwell Acioli, Arthur Moreno, Victor Holanda e Fabiano Conrado, que contribuíram para o desenvolvimento acadêmico, ao compartilhar experiência e tempo durante a elaboração deste trabalho.

Ao meu orientador Prof. Dr. Rafael de Amorim Silva e co-orientador Prof. Dr. Bruno de Almeida Pimentel, pelo suporte, disponibilidade e orientações, sempre fomentando a busca pela excelência, e no apontamento dos caminhos a serem percorridos.

Aos professores Drs. Almir Pereira Guimarães e Diego Demerval Medeiros da Cunha Matos, pela presença na banca de defesa.

A todos os professores do Programa de Pós Graduação em Informática pelo compartilhamento dos conhecimentos ao longo desses anos de curso.

Gratidão por me acompanharem nessa caminhada.

"O desejo profundo da humanidade pelo conhecimento é justificativa suficiente para nossa busca contínua. E nossa meta não é nada menos do que uma descrição completa do universo onde vivemos."

– Hawking, Stephen.

Resumo

A tecnologia utilizada na medicina evolui a cada momento, e com isso novas soluções surgem. Com a inteligência artificial e aprendizagem de máquina aplicada em problemas da área médica, a capacidade de antecipação à diagnósticos e prognósticos aumentaram consideravelmente, devido a capacidade de processamento de dados que um computador pode realizar. Foi aplicado no presente trabalho algoritmos de aprendizagem de máquina, a fim de obter prognóstico de uma doença conhecida na medicina, a SEPSES, que causa uma infecção generalizada. Septicemia é uma infecção que se não identificada corretamente, compromete significativamente a saúde dos órgão próximos, e pode causar falência múltipla dos órgãos, desse modo é necessário adotar medidas preditivas para poder conter o avanço da doença. Portanto, a metodologia deste trabalho consistiu em apresentar um modelo supervisionado para o prognóstico de pacientes que possuem sintomas de Sepses. Por fim, a proposta integrou a elaboração de modelos de aprendizagem de máquina no intuito de obter uma predição no prognóstico de SEPSE a partir de dados coletados no monitoramento dos pacientes internados em UTI por meio de dados disponibilizados no programa PhysioNet - eICU Collaborative Research Database. Com isso, de acordo com os modelos implementados, regressão logística, K-vizinhos mais próximos (KNN) e Máquina de Vetores de Suporte. O melhor resultado de cada modelo entre os cenários explorados para prognóstico de mortalidade, de acordo com AUC do método estatístico, característica de operação do receptor (ROC), foi de 0.82, 0.74 e 0.85, respectivamente. E para a predição do tempo de internação do paciente, de acordo com o melhor cenário, os modelos obtiveram AUC 0.85, 0.64 e 0.32, respectivamente.

Keywords: prognóstico, diagnóstico, Sepses, UTI e Aprendizado de Máquina.

Abstract

The technology used in medicine evolves all the time, and with that new solutions emerge. With artificial intelligence and machine learning applied to medical problems, the ability to anticipate diagnoses and prognoses can be considered, due to the data processing capacity that a computer can perform. No machine learning problem was applied, a prognostic end of a disease known in medicine, SEPSIS, which causes a generalized infection. Sepsis is an effective prevention of measures not correctly identified, significantly compromising the destruction of Organs following organs, and can cause the destruction of multiple organs, so it is necessary to adopt measures to be able to contain the advance of pre-health. Therefore, the methodology of this work is to present a model of supervision for the prognosis of patients who present symptoms of sepsis. Finally, a proposal integrated the elaboration of learning models in order to obtain a prediction from an ICU prediction data program in the monitoring of inpatients from PhysioNet S in the monitoring of inpatients from ICU research data - eICU Collaborative Database through data used in the research database. With that, with the models vector models according to the solution, K-nearest neighbors (KNN) and Support machines. The best result of receiver operation (ROC) was according to the method of each model (ROC), respectively. And for the prediction of patient time, according to the best scenario, the models obtained AUC 0.85, 0.64 and 0.32, respectively.

Keywords: prognostics, diagnostics, Sepsis, UTI and Machine Learning.

Sumário

Lista de Figuras	vii
Lista de Tabelas	viii
1 Introdução	12
1.0.1 Estrutura	14
2 Septicemia	15
2.0.1 Diagnóstico	15
2.1 Septicemia	16
2.1.1 SIRS	18
2.1.2 SOFA	18
2.1.3 qSOFA	19
2.2 Prognóstico	21
2.2.1 Dados de pacientes no Brasil	22
3 Fundamentação	25
3.1 Domínio de Aplicação	25
4 Relato do Problema	28
4.1 Definição	28
4.2 SAPS-3	28
4.3 APACHE IV	29
4.4 Trabalhos Relacionados	30
4.4.1 Diagnóstico	30
4.4.2 Prognóstico	32
5 Proposta	34
5.1 Fundamentação	34
5.1.1 Fluxo de normalização dos dados	35
5.2 Hipóteses	36
6 Metodologia de Experimento	38
6.1 Métodos	38
7 Validação	40
7.1 Ferramentas	40
7.2 Análise dos resultados	40
7.2.1 Característica de Operação do Receptor - curva ROC	40
7.3 Cenário	41
7.4 Metodologia	43

7.4.1	KNN - K-Nearest Neighbor	44
7.4.2	RL - Regressão Logística	44
7.4.3	SVM - Máquina de Vetores de Suporte	45
7.5	Parâmetros	46
7.6	Experimentos	46
8	Resultados	48
8.1	Cenários	48
8.1.1	Cenário 1	49
8.1.2	Regressão Logística	49
8.1.3	KNN	49
8.1.4	SVM	50
8.2	Cenário 2	51
8.2.1	Regressão Logística	51
8.2.2	KNN	52
8.2.3	SVM	53
8.3	Cenário 3	53
8.3.1	Regressão Logística	54
8.3.2	KNN	54
8.3.3	SVM	56
8.4	Cenário 4	56
8.4.1	Regressão Logística	56
8.4.2	KNN	56
8.4.3	SVM	57
9	Discussão	59
10	Conclusão	61
	Referências	62
A	Código para gerar os modelos de aprendizagem de máquina	68

Lista de Figuras

2.1	Tipos de Sepsis	17
5.1	Resultado da calculadora web APACHE IV ICU Calculators	35
5.2	Fluxo aplicado no desenvolvimento dos modelos	36
7.1	Exemplo Curva ROC	41
7.2	Exemplo de Hiperplano - scikit-learn (2022)	45
8.1	Matriz de Confusão Regressão Logística - cenário 1	49
8.2	Matriz de Confusão KNN - cenário 1	50
8.3	Matriz de Confusão SVM - cenário 1	50
8.4	Curva ROC dos modelos - cenário 1	51
8.5	Matriz de Confusão Regressão Logística - cenário 2	51
8.6	Curva ROC dos modelos - cenário 2	52
8.7	Matriz de confusão KNN - cenário 2	52
8.8	Matriz de Confusão SVM - cenário 2	53
8.9	Matriz de confusão Regressão logística - cenário 3	54
8.10	Matriz de Confusão KNN - cenário 3	55
8.11	Matriz de Confusão SVM - cenário 3	55
8.12	Curva ROC dos modelos implementados -cenário 3	55
8.13	Matriz de confusão Regressão Logística -cenário 4	56
8.14	Matriz de confusão KNN -cenário 4	57
8.15	Matriz de confusão SVM -cenário 4	57
8.16	Curva ROC dos modelos implementados -cenário 4	58

Lista de Tabelas

1.1	Evolução na definição da SEPSE	14
2.1	Critério Síndrome de resposta inflamatória sistêmica - SIRS	18
2.2	falha sequencial de órgãos - SOFA - adaptado de Jones et al. (2009)	19
2.3	Rápida Avaliação da pontuação na falha sequencial de órgãos - qSOFA - adaptado de Marik and Taeb (2017)	20
2.4	Quantidade de Internacoes por Septicemia - Dezembro 2018 até Marco 2019	23
2.5	Quantidade de Internacoes por Septicemia - Dezembro 2019 até Marco 2020	23
2.6	Quantidade Internacoes por Septicemia - Dezembro 2020 até Marco 2021	23
2.7	Quantidade Internacoes por Septicemia - Dezembro 2021 até Marco 2022	24
2.8	Mortalidade por Septicemia - Janeiro 2022 até Junho 2022	24
2.9	Número médio de internações por Septicemia - Anos 2018,2019,2020 e 2021	24
4.1	Taxa de mortalidade pelo SAPS e Tempo de internação. Adaptado de Zimmerman et al. (2006)	30
4.2	Comparação dos algoritmos dos trabalhos quanto a prognóstico de Mortalidade.	33
7.1	Tabelas Disponíveis com dados clínicos dos pacientes.	42
7.2	Variáveis da base de dados do eICU Brooks and Tsao	43
7.3	Característica dos pacientes internados presente na base de dados.	43
7.4	Parâmetros utilizados nos modelos após normalização dos dados.	46

1

Introdução

Septicemia ou Sepsé é uma doença infecciosa generalizada, mais precisamente uma infecção sanguínea. Segundo o Instituto Latino americano de sepsés (ILAS), pacientes com septicemia ocupam no Brasil cerca de 25% dos leitos de UTI - Unidade de Terapia Intensiva-, além de ser a principal causa de morte nos hospitais do Brasil, com cerca de 670 mil mortes anuais.

[Ren et al. \(2020\)](#) relataram a correlação entre pessoas com septicemia e o vírus SARS-COV2, ou COVID-19, devido a semelhança dos sintomas entre as doenças, principalmente quanto ao agravamento do estado do paciente na UTI. O artigo relata que há uma alta probabilidade do paciente que está diagnosticado com COVID-19, desenvolver a infecção por SEPSE por agravamento dos sintomas causados pelo coronavírus.

Além disso, os pesquisadores [Li et al. \(2020\)](#) também observaram, que os pacientes apresentavam suspeita de sepsés e choque séptico, em pacientes diagnosticados com COVID-19, de acordo com a definição de SEPSES-3, como está presente na tabela 1.1.

Dados da organização mundial da saúde (OMS), as principais características de pessoas que desenvolvem essa infecção, são pessoas que vivem em locais desprovidos de recursos, isto é, como saneamento e a educação sanitária. Recém-nascidos e gestantes representam cerca de 85% dos casos registrados no mundo. Há também uma preocupação acerca da acessibilidade prévia quanto a cuidados básicos da saúde,

A motivação deste trabalho, parte do direcionamento dos estudos da área da informática, para auxiliar a área médica com tecnologias cada vez mais robustas. Com isso, ser capaz de automatizar processos e realizar funções complexas de maneira mais rápida, que só a computação é capaz de realizar, e assim contribuir na tomada de decisão dos especialistas, de forma mais assertiva.

Desse modo, o trabalho tem o foco no desenvolvimento de um modelo para prognóstico de pacientes internados na UTI com SEPTICEMIA. E assim será possível utilizar os modelos de aprendizado de máquina, no auxílio das tomadas de decisões.

A taxa de mortalidade pela Sepse é a maior já registrada em hospitais [L. et al. \(2021\)](#), além disso, estudos como o de [Lobo et al. \(2019\)](#) relataram que é a doença que tem uma média de 15 a 17 milhões de pacientes anuais, com o índice de mortalidade anual de 5 milhões. No Brasil só em 2019, foram registrados 21671 mortes relacionados a Sepse.

É importante ressaltar, que a infecção em si, não está associada a contaminação geral de órgãos, mas na maioria dos casos trata-se do paciente estar contaminado pela bactéria *Estreptococcias*, e a partir disso, transmite a infecção à um órgão específico. Após essa infecção, é possível atingir outros órgãos indiretamente, por isso, têm-se o risco eminente de uma infecção generalizada, que tende a comprometer os órgãos afetados, assim, com a evolução máxima da doença, os órgãos falharão.

Além da taxa de mortalidade por esta doença ser alta, segundo o trabalhos dos pesquisadores [Seymour et al. \(2017\)](#), a cada hora de atraso na conclusão da coleta do pacote de 3h no diagnóstico, e administração dos dados clínicos dos pacientes, as chances de mortalidade do paciente se agrava em 4%.

O monitoramento instaurado na UTI, se deve as complicações causadas pela infecção, com isso os hospitais levam tempo para identificar, não só o tipo de sepses, como também o melhor tratamento para a doença. Isso ocorre nos casos em que é identificado corretamente o diagnóstico, e por isso o gasto nos hospitais com esta doença se torna bastante elevado.

Devido a quantidade de casos de pessoas acometidas por essa doença [Zonta et al. \(2018\)](#), outro fator determinante para a internação das pessoas, é que a manifestação dos sintomas, podem ser confundidos com outras doenças. Portanto, geram falsos positivos, ao tentar predizer antecipadamente o diagnóstico, como também o prognóstico desses pacientes.

Há vários tipos de Sepses definidas pelo CID-10 ¹, -ver figura 2.1. Além dessa classificação, há também a nova definição para as sepses, a SEPSES-3. [Jordão et al. \(2019\)](#) relataram a evolução pela qual a definição de SEPSES passou ao longo dos anos, e teve mudanças desde as definições da SEPSES-01 e SEPSES-2 - ver tabela 1.1-, com isso, o diagnóstico específico e preciso, fica comprometido quanto a acurácia na classificação da doença.

Como foi visto anteriormente, quanto mais tempo demora para o prognóstico e o tratamento da Sepse, a chance de contaminação dos outros órgãos aumenta. Portanto, o diagnóstico quanto mais exato e rápido, melhor para o prognóstico da infecção. Principalmente para evitar o choque séptico, este estado clínico, faz com que os órgãos afetados, tenham mal funcionamento, de modo temporário, ou permanente, caso o paciente venha a sobreviver.

A proposta do trabalho é avaliar os casos de prognóstico dos pacientes acometidos com septicemia, a fim de melhorar os resultados já obtidos previamente com a utilização das técnicas APACHE (Acute Physiology and Chronic Health Evaluation) e SAPS (Simplified Acute Physiology Score), e utilizar os resultados como variáveis nos algoritmos propostos, para prognóstico de tempo de permanência na UTI e de mortalidade.

¹CID-10 -Classificação Internacional de Doenças

SEPSE	CHOQUE SÉPTICO
Suspeita/Documentação de infecção + Aumento de 2 ou mais pontuação SOFA	Sepses + Necessidade de vasopressores para manter Pressão arterial Média > 65 Lactato > 2mmol/L após reanimação volêmica adequada.

Tabela 1.1: Evolução na definição da SEPSE

1.0.1 Estrutura

A estrutura do trabalho está composta pelo capítulo 2 - Prognóstico-, numa abordagem de como é realizado hoje os prognóstico de pacientes, especificamente o prognóstico de mortalidade e suas técnicas e critérios utilizados na medicina atual, segue assim no capítulo 3 -Fundamentação- com a descrição do domínio de aplicação deste trabalho, em seguida no capítulo 4 - Relato do Problema - que descreve o problema de pesquisa do prognóstico de pacientes acometidos com sepses, e correlaciona os métodos APACHE e SAPS. O trabalho segue com o capítulo 5- Proposta - que descreve detalhadamente a proposta deste trabalho, o caminho percorrido dentro da revisão da literatura até o porquê deste problema de pesquisa. Por fim, temos os capítulos 6 e 7, que dispõem da validação e os resultados, respectivamente, para concluir com os resultados e discutir posteriormente nas seções 8 e 9 o trabalho realizado.

Capítulo 2

Septicemia

Este capítulo descreve os métodos aplicados para diagnosticar e prognosticar atualmente, e também os dados de internação e mortalidade pela doença, no Brasil.

2.0.1 Diagnóstico

No âmbito hospitalar, para o diagnóstico da sepsis são avaliados os sintomas do paciente ao ser internado na unidade de terapia intensiva (UTI). Além disso, de acordo com o CID-10 a depender do tipo de sepsis identificada, o paciente terá um prognóstico diferente.

Desse modo, uma das características mais importantes para o diagnóstico, é a presença da bactéria *Estreptococcus* no indivíduo. É válido informar que há uma predominância maior em pacientes do neonatal, uma vez que há uma incidência maior em gestantes. Sendo assim, o micróbio pode ter se desenvolvido na corrente sanguínea, como também na região genital da paciente, o que potencializa as chances do recém nascido estar mais sujeito à contaminação.

Isto posto, além das características citadas, ainda há outro modo de contaminação o qual ocorre por meio de contatos entre humanos e animais. Santos (1999), observaram que a propagação ocorre mediante a apertos de mãos, compartilhamento de objetos pessoais ou até mesmo pelo ar (caso algum indivíduo tenha transmitido ao ambiente através de espirro). Por conseguinte, o autor ressalta o cuidado necessário com o contato direto com sangue contaminado, uma vez a infecção também é propagada por meio de feridas expostas.

Portanto, é importante salientar que uma das principais causas para a colonização e evolução da bactéria *Estreptococcus* está relacionada à parte higiênica pessoal. Embora a situação pandêmica, causada pela COVID-19, tenha contribuído para a educação das pessoas no que diz respeito às boas práticas de higiene, sabe-se que ainda existem diversas lacunas. Desse modo, espera-se que o cuidado aumente quanto à higiene, o que diminuirá os riscos da propagação de doenças infecciosas.

2.1 Septicemia

Nesta seção serão abordadas as características que definem a sepse, e quais os principais sintomas a serem avaliados no diagnóstico. Também será descrito quais os meios que os profissionais utilizam detalhadamente, para então chegar em determinado diagnóstico.

A medicina passou por diversas transformações nas últimas décadas, isso se deve e muito não só ao fator tecnológico, mas também das pesquisas laboratoriais que são realizadas com critérios e procedimentos que são seguidos fielmente pela comunidade médica.

Quanto a evolução do diagnóstico de sepse se deu desde o critério elaborado e definido por [Bone et al. \(1992\)](#), chamado de Síndrome da resposta inflamatória sistêmica (SIRS), que têm em sua definição pela SEPSES-1, critérios de avaliação que validavam, se o paciente possuía sepse ou não. Alguns estudos como [Marik and Taeb \(2017\)](#), propuseram uma análise desse método, no qual avalia as respostas conforme o monitoramento do paciente.

Com o passar do tempo, para avaliar a eficiência do modelo, estudiosos chegaram a conclusão de que esse critério não era suficiente para o diagnóstico da sepse, mas somente funcionava para uma classificação de suspeita de infecção. A partir disso, propuseram que os pacientes realizassem mais exames com o fito de detectar outras características a nível laboratorial para apontar se haviam chances ou não do indivíduo estar contaminado pela sepse.

Com os critérios realizados na UTI, fica evidente a quantidade de variáveis que os médicos e enfermeiros devem lidar, principalmente por se tratar de análise da semiologia, ao se tratar de ter avaliação visual e tátil em alguns dos sintomas. O encaminhamento para análise laboratorial prejudica a curto prazo o tratamento efetivo do paciente, com a necessidade cada vez maior de poder detectar da maneira mais rápida possível a infecção generalizada.

Isto posto, antes da cultura do sangue do indivíduo, e esperar os dados laboratoriais, é observado outros sintomas que já foram mencionados, como temperatura corporal, se possui alguma inflamação nas vias respiratórias, batimentos cardíacos ou outros sintomas que o paciente possa relatar.

O coleta e examinação do sangue, se deve a obter uma maior acurácia quanto ao diagnóstico, devido a probabilidade de ser um falso positivo ser alta, pois, os sintomas são comuns para outros tipos de doenças.

Um paciente com suspeita de sepse pode apresentar irritação na garganta, tosse forte, temperatura corporal maior que 38°C ou menor que 36°C, além de batimentos cardíacos acima de 90 por minuto, de acordo com o critério da síndrome de resposta inflamatória sistêmica, ou SIRS, [Kim and Park \(2019\)](#) o paciente também apresenta uma respiração por minuto maior que 20 e pressão parcial de CO₂ menor que 32mm Hg.

Além disso, há também o critério de avaliação de falência de órgãos sequencial (SOFA) [Jones et al. \(2009\)](#), que é um sistema de pontos que são utilizados para determinar se o paciente possui SEPSE. No entanto, este método apesar de ser bastante eficiente, não atende as expectativas ao se tratar de diagnóstico rápido, principalmente porque depende de dados

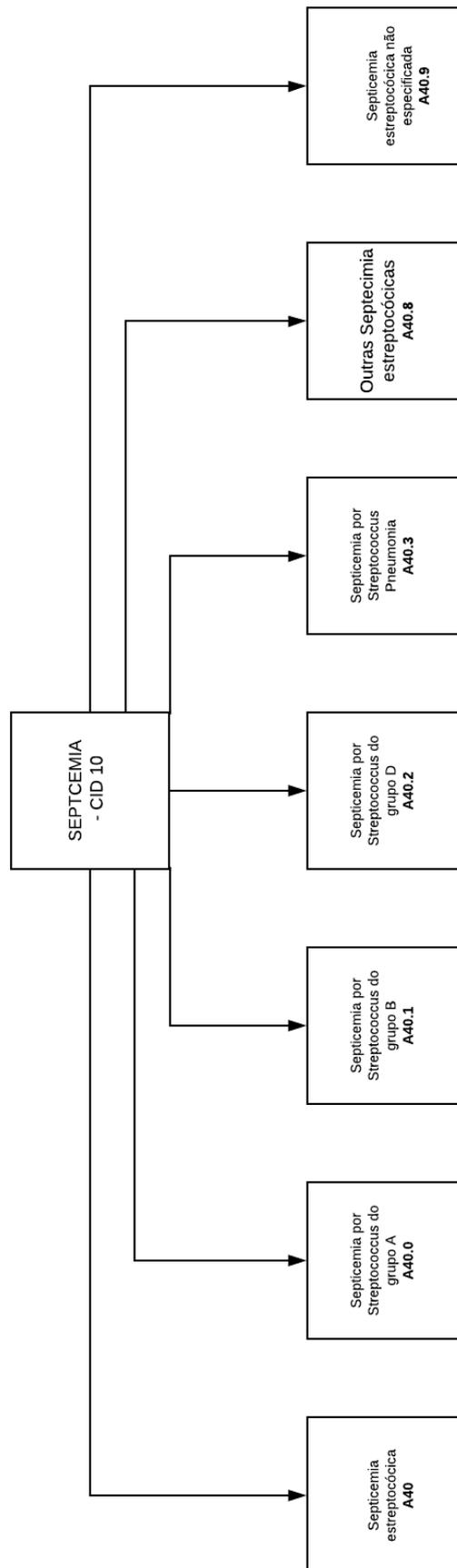


Figura 2.1: Tipos de Sepse

laboratoriais.

Isto posto, devido a necessidade e urgências, foi desenvolvido outro método baseando-se no SOFA, o qSOFA, uma metodologia mais rápida, que determina que um paciente que obtém pontuação maior ou igual a 2, possui falência de órgão, conseqüentemente é diagnosticado com sepse severa. Isso é devido ao monitoramento do paciente que aparentar pressão sanguínea sistólica menor ou igual a 100 mm Hg. Por fim, têm-se a avaliação da escala de coma de Gasglow, que observa alterações no estado mental do paciente, onde, se for menor que 15, de acordo com a tabela de ¹, o paciente possui falência de órgãos, e tem maior probabilidade de permanência na UTI, conseqüentemente maior mortalidade [Jordão et al. \(2019\)](#).

2.1.1 SIRS

As características avaliadas estão presentes na tabela 2.1, logo abaixo, e é possível analisar quais os sintomas que o paciente apresenta na UTI. Este método é bastante eficiente quanto a classificar o paciente, porém, não é recomendado para um diagnóstico preciso, isso se deve ao índice considerável de falsos positivos, a presença em pacientes saudáveis e em outros pacientes com doenças que não caracteriza SEPSE.

Sintomas	Parâmetros
Taquicardia	Batimentos acima de 90 por minuto
Febre ou Hipotermia	Temperatura maior que 38°C ou menor que 36°C
Células de defesa (células brancas)	Presença no sangue menor que 1,200/mm ³ , maior que 4,000/mm ³ ou bandemia maior ou igual a 10%

Tabela 2.1: Critério Síndrome de resposta inflamatória sistêmica - SIRS

2.1.2 SOFA

O resultado desse método, irá variar de hospital para hospital, isso depende do quão rápido será a análise laboratorial, que identifica as variáveis que estão presentes na tabela ??.

Este critério é o mais longo e mais detalhado que pode ser executado, e com ele é possível diagnosticar o paciente, isso é possível devido a coleta do sangue e análise laboratorial detalhada.

¹Paul M. Brennan

Pontuação				
Sintoma	1	2	3	4
Respiração				
PaO ₂ /FIO ₂ (mm Hg)	<400	<300	<220	<100
SaO ₂ /FIO ₂	221-301	142-220	67-141	<67
Coagulação				
Platelets×10 ³ /mm ³	<150	<100	<50	<20
Fígado				
Bilirubin (mg/dL)	1.2-1.9	2.0-5.9	6.0-11.9	>12.0
Cardiovascular				
Hypotension	MAP <70	Dopamine 5 or dobutamine (any)	Dopamine >5 or norepinephrine 0.1	Dopamine >15 or norepinephrine >0.1
Sistema Nervoso Central				
Glasgow Coma Score	13-14	10-12	6-9	<6
Renal				
Creatinine (mg/dL) or urine output (mL/d)	1.2-1.9	2.0-3.4	3.5-4.9 or <500	>5.0 or <200

Tabela 2.2: falha sequencial de órgãos - SOFA - adaptado de [Jones et al. \(2009\)](#)

2.1.3 qSOFA

O método qSOFA foi desenvolvido a partir do método SOFA, os critérios são coletado mais rapidamente. Com isso, é possível realizar uma triagem do paciente e um possível redirecionamento para outros exames. Este sistema possui uma eficiência considerável na predição da mortalidade, mas não é recomendável para o diagnóstico preciso da SEPSE.

Nessa perspectiva, estudos também demonstraram que, na comparação entre o critério SIRS e qSOFA, na predição do diagnóstico de doença infecciosa em pacientes, o sistema SIRS superou o qSOFA no quesito predição. A partir disso, a análise ocorreu em pacientes com suspeitas de infecção com a especificidade e sensibilidade, mesmo que o qSOFA não tenha obtido um desempenho melhor na predição do diagnóstico, o sistema tem maior eficiência quanto a predição da mortalidade dos pacientes [Marik and Taeb \(2017\)](#).

Desse modo, alguns autores analisaram o desenvolvimentos desses modelos ao longo dos anos, e alguns fatores determinantes fizeram com que considerassem o método SIRS, ser descontinuado para o diagnóstico de sepse. O sistema possui uma tendência quanto aos resultados para falsos positivos, como também os sintomas foram observados em pacientes saudáveis [Marik and Taeb \(2017\)](#). As pontuações e as características deste sistema estão presentes nas tabelas a seguir 2.3 .

Essa descontinuidade fica mais evidente, quando se é comparado a evolução na definição

dos quadros clínicos dos pacientes [Kim and Park \(2019\)](#), SEPSE-1, SEPSE-2 e SEPSE-3 atual definição 1.1.

Sintomas Avaliados	Parâmetros	Pontuação
Taxa de Respiração	$\geq 22/\text{min}$	1
Mudança no Estado Mental	sim	1
Pressão sistólica do Sangue	$\leq 100 \text{ mmHg}$	1

Tabela 2.3: Rápida Avaliação da pontuação na falha sequencial de órgãos - qSOFA - adaptado de [Marik and Taeb \(2017\)](#)

A utilização de sistemas ou frameworks para avaliação dos pacientes, se faz necessário no âmbito hospitalar, pois, facilita e otimiza o tempo de registro do especialista, ao realizar o exame clínico. Dentre os sistemas citados, o SIRS não têm validação suficiente, para definir se o paciente possui ou não sepse, uma vez que ele tem resposta em pacientes em condições benignas, como também o qSOFA que auxilia na pontuação do SOFA, e não pode ser utilizado para diagnosticar se há ou não infecção por sepse no paciente, e sim para conduzir um melhor prognóstico [Jordão et al. \(2019\)](#).

Com o problema de antecipar diagnóstico [Zonta et al. \(2018\)](#) para o melhor tratamento e cura do paciente, o desenvolvimento contínuo de métodos e aplicação desses critérios, são cada vez mais necessários, pois a incidência dessa infecção nas UTI's cresce continuamente, devido as diversas variantes que podem ocasionar a SEPSES, dentre elas vírus, bactérias, fungos e outras doenças que podem agravar a situação do paciente e possibilitar a infecção.

Nesse interím, diante da dificuldade em elaborar novos métodos para uma predição antecipada de doenças, o envolvimento de trabalhos na literatura ligados à computação contribuíram cada vez mais na obtenção de resultados mais rápidos.

No quesito avanço tecnológico, a velocidade computacional ganhou reconhecimento, principalmente nos setores industriais, desde a revolução industrial, com seus processos manufaturados. Estes processos, foram propulsores no desenvolvimento de novas tecnologias, na busca por desenvolver cada vez mais materiais o poder computacional com as máquinas.

Esse desenvolvimento tecnológico, culminou em trabalhos que efetuam cálculos, análises, predições, entre outros métodos explorados pela ciência, que obtém resultados mais rápidos, com precisão melhor ou igual a avaliação humana, até mesmo problemas que nenhum humano teria capacidade de resolver, intelectual ou fisicamente. Isso ocorre devido ao poder de processamento dos computadores atuais e aumentam cada vez mais o nível de confiança da população quanto ao uso de novas tecnologias.

Dessa forma, a metodologia ocorre da seguinte forma: o prognóstico de pacientes normalmente é dado por scores, que são calculados minuciosamente por frameworks que indicam um percentual, seja para definir o tipo de tratamento a ser seguido pelos especialistas com os pacientes, seja para indicar o índice de mortalidade dos pacientes.

2.2 Prognóstico

Esta seção descreve as metodologias APACHE para prognóstico de mortalidade de pacientes e a metodologia SAPS, mais precisamente os métodos mais atuais, o Apache IV e SAPS-3, que são métodos responsáveis pelo prognóstico de mortalidade de pacientes acometidos na UTI, e evidencia o uso mundial destes frameworks, que auxiliam através de uma análise de risco com os sistemas desenvolvidos pelos autores.

O estudo de avaliação de riscos em pacientes presentes na UTI, foi iniciado por Knaus em 1981 [Knaus et al. \(1981\)](#), no desenvolvimento do sistema APACHE (Acute Physioly and Cronic Health Evaluation), para avaliar o nível de gravidade de doença em pacientes. Embora este sistema esteja defasado, o mesmo grupo de estudo conseguiu otimizar o sistema ao diminuir o número de variáveis de 37 para 17, com uma nova denominação do sistema, o APACHE 2, com variação na pontuação de 0 até 71 de acordo com [JE \(1985\)](#).

Com o desenvolvimento do método APACHE ao longo dos anos, foi possível otimizar por mais três vezes o método, para o APACHE III [Knaus et al. \(1991\)](#), IV [Zimmerman et al. \(2006\)](#). Assim, com o métodos IV, ao observar último método de prognóstico, foi possível obter uma base de dados, disponível pelos autores [Balkan et al. \(2018\)](#), no qual é fornecido resultados da pontuação APACHE IV.

Por fim, através da análise desses dados, é possível obter uma validação do modelo a ser criado por este trabalho. O banco de dados utilizado, contém dados clínicos e administrativos de mais de 400 locais hospitalares que empregam tecnologia de telemedicina nos Estados Unidos entre 2014 e 2015. Os detalhes presentes nos dados recolhidos pelos especialistas, permitem a análise retrospectiva de dados, e a avaliação das ferramentas existentes de apoio à decisão clínica, bem como o desenvolvimento de novos modelos analíticos.

Ademais, na avaliação de prognóstico há alguns frameworks disponibilizados na internet, esses frameworks calculam as entradas dos usuários: seja médico, enfermeiro ou algum estudante que queira simular determinados dados de pacientes, e assim obter uma previsão do prognóstico de mortalidade do paciente. Com a utilização desses sistemas, é possível obter a pontuação do paciente para avaliar a gravidade da doença. Por fim, ter uma tomada de decisão melhor, por exemplo em pacientes no pré ou pós operatório, a depender do grau de risco.

No que diz respeito ao entendimento da área médica, para uma análise de diagnóstico e prognóstico, a semiologia é aplicada na identificação dos sinais e sintomas expressados pelo paciente, sejam os sintomas expressados verbalmente ou sinais não verbais, dos quais os médicos e enfermeiros estão aptos a identificar no período de internação do paciente na UTI. Abaixo, além de demonstrar os sinais e sintomas avaliados, é possível observar a evolução do sistema APACHE para prognóstico de mortalidade.

Frameworks

O sistema APACHE foi desenvolvido e aprimorado por [Knaus et al. \(1981\)](#), a fim de observar de maneira enviesada os pacientes presentes nos leitos de UTI, conseguiram diminuir a quantidade de variáveis observadas, incluí-se o próprio pontuação do APACHE original.

Elaborado por Knaus 1981 e simplificada em 1985. O método requer confirmação do sistema que melhor classifica o paciente de acordo com a doença, o SIRS. O objetivo do método é o prognóstico de doenças agressivas, principalmente pessoas adultas que deram entrada na UTI previamente nas ultimas 24h.

É recomendado pelo autor que seja realizada o teste da SIRS, para utilizar o método em pacientes com doenças infecciosas (septicemia) e poder predizer a mortalidade. Consequentemente obter a probabilidade de recuperação do paciente.

Desse modo, alguns artigos não recomendam a utilização desse método devido a alguns estudos que o apache 2 foi comprovado enquanto ineficaz. Já a maioria dos estudos apontam a capacidade de predizer a mortalidade do paciente de acordo com as variáveis. Assim, no Brasil foi aprovado este método estadunidense pela Portaria 3432 ².

Apesar do diagnóstico precoce ser um fator fundamental na taxa de sucesso de sobrevivência do paciente, o trabalho é voltado para o prognóstico de pacientes, adultos que foram diagnosticados com septicemia, e possuem um tempo de internação igual ou superior a 24 horas na UTI.

Dessa forma, na avaliação de risco do paciente, no cenário atual, há dois frameworks que são capazes de obter resultados com nível de confiança aceitável, são eles: APACHE IV, desenvolvido após uma evolução do sistema, desde [Knaus et al. \(1981\)](#) com o APACHE, por [Zimmerman et al. \(2006\)](#), e o APS (Acute Physiology Score) baseado no APACHE também de [Knaus](#), os pesquisadores [Moreno et al. \(2005\)](#) trabalharam no desenvolvendo SAPS (Simplified Acute Physiology Score), e conforme evoluíram na pesquisa, utilizam hoje o SAPS 3, com 20 variáveis no seu modelo.

2.2.1 Dados de pacientes no Brasil

No Ministério da Saúde - Sistema de Informações Hospitalares do SUS (SIH/SUS), no site [Ministério-da-Saúde \(2021\)](#). É possível extrair dados importantes sobre internações nos últimos anos. Informações com relação as internações que ocorreram no período de dezembro de 2020 e março de 2021, a quantidade de internações média por septicemia foi de 9280 internações em todas as regiões do Brasil, de acordo com a tabela 2.6.

Com isso, Para obter uma comparação quantitativa dos períodos, principalmente os períodos pré e pós pandemia da covid-19, utilizou-se da plataforma, para obtenção dos dados referente as internações nesse período, números presente na tabela 2.4 e 2.5.

²Portaria 3432, de 12 de agosto de 1998(9) - 3.2. Avaliação através do APACHE II se for UTI Adulto, o PRISM II se UTI Pediátrica e o PSI modificado se UTI Neonatal.

Logo depois, no acompanhamento das internações hospitalares no Brasil, foi obtido através a base de dados presente no site [Ministério-da-Saúde \(2021\)](#), mais dados quanto ao número de internações. Por fim, foi avaliado o mesmo período das avaliações anteriores, como também as mesmas regiões, e está contido na tabela 2.7.

Além dessas avaliações quanto ao número de internações, foi obtido na mesma base de dados, desde Janeiro de 2022 até Junho de 2022, números correspondentes a taxa de mortalidade por sepses, nas regiões do Brasil, conforme tabela 2.8.

Nesse intervalo, destaca-se a média de internações de pacientes acometidos por septicemia no Brasil. Com isso, os números presentes na tabela 2.9, remetem o aumento de pessoas internadas ao longo dos anos. Exceto a região Norte, as demais regiões tiveram um aumento quanto a média de pessoas internadas, ao considerar o retrospecto mensal de acordo com o mês de Dezembro.

Regiao	Dez	Jan	Fev	Mar
Norte	579	615	560	515
Nordeste	1966	2126	1943	2158
Sudeste	5486	6020	5524	5860
Sul	2045	2345	2089	2151
Centro-Oeste	400	529	565	547

Tabela 2.4: Quantidade de Internacoes por Septicemia - Dezembro 2018 até Marco 2019

Regiao	Dez	Jan	Fev	Mar
Norte	592	608	583	582
Nordeste	2305	2432	2281	2309
Sudeste	5757	6477	5713	6063
Sul	2133	2497	2445	2395
Centro-Oeste	580	615	597	613

Tabela 2.5: Quantidade de Internacoes por Septicemia - Dezembro 2019 até Marco 2020

Regiao	Dez	Jan	Fev	Mar
Norte	464	455	506	374
Nordeste	1670	1809	1859	1635
Sudeste	4831	5123	4969	4679
Sul	1750	1755	1748	1554
Centro-Oeste	511	535	424	472

Tabela 2.6: Quantidade Internacoes por Septicemia - Dezembro 2020 até Marco 2021

Regiao	Dez	Jan	Fev	Mar
Norte	763	728	745	761
Nordeste	2.362	2.325	2.084	2408
Sudeste	6.002	6.567	5.616	6.128
Sul	2.080	2.224	2.238	2.356
Centro-Oeste	510	696	560	624

Tabela 2.7: Quantidade Internacoes por Septicemia - Dezembro 2021 até Marco 2022

Regiao	Jan	Fev	Mar	Abr	Mai	Jun
Norte	42,03	38,93	43,36	39,82	38,63	37,15
Nordeste	47,53	47,84	44,52	47,20	43,88	46,44
Sudeste	50,80	50,62	48,25	48,46	48,6	50,14
Sul	43,44	44,28	39,86	41,35	43,60	41,86
Centro-Oeste	35,06	35,54	37,66	35,12	35,25	39,30

Tabela 2.8: Mortalidade por Septicemia - Janeiro 2022 até Junho 2022

Regiões	Período				
	2018	2019	2020	2021	junho 2022
Norte	3.636,71	3.722,39	3.980,50	3.649,39	3.638,51
Nordeste	3.578,42	3.664,55	3.948,00	4.092,91	4.519,05
Sudeste	3.848,12	3.639,86	4.265,00	4.268,61	4.771,46
Sul	3.502,12	3.582,44	3.985,80	3.798,03	4.173,10
Centro-Oeste	3.317,75	3.796,55	3.342,99	3.827,56	5.341,97

Tabela 2.9: Número médio de internações por Septicemia - Anos 2018,2019,2020 e 2021

Capítulo 3

Fundamentação

Neste capítulo está descrito toda a base de conhecimento agregado neste trabalho, a partir dos domínios de aplicação, que foram adquiridos ao longo do desenvolvimento do mesmo, e implementados de acordo com o problema de pesquisa proposto.

3.1 Domínio de Aplicação

Com o desenvolvimento tecnológico existente, principalmente ao se tratar de computadores, é possível a implementação de sistemas que possam automatizar metodologias e aplicações. Devido a alta capacidade de processamento computacional, é possível fazer análise de grandes quantidades de dados. O avanço tecnológico permitiu que cientistas conseguissem produzir hardwares cada vez mais robustos, para otimizar o tempo de leitura e gravação de dados.

A ciência de dados é a área de estudo que lida com análise e tratamento de arquivos provenientes de um sistema. As informações são tratadas, independente da quantidade de dados existentes, isso dependerá do modo em que é implementado a análise. Essa área contempla também a organização e implementação de algoritmos de aprendizado de máquina, que são capazes de realizar plotagens gráficas, filtros avançados, comparativos, classificação e predição.

Desse modo, Todas essas funcionalidades podem ser utilizadas de acordo com o algoritmo que for implementado pelo desenvolvedor, o que possibilita o desenvolvimento de programas complexos para a automatização e otimização de um determinado problema.

Além disso, atualmente, com a robustez dos hardwares atuais, o desenvolvimento de técnicas de programação avançaram tanto, os cientistas, engenheiros e programadores conseguem implementar algoritmos de inteligência artificial com um nível de complexidade maior e bem mais assertivo.

Outrossim, a consequência desses avanços, se deve aos problemas científicos, e assim têm-se o intuito, de prever acontecimentos em tempos mínimos, em questão de segundos, ou milésimos de segundos, que ao ser comparado com análise humana, demandaria muito mais

tempo para obter resultados significativos.

Isto posto, os algoritmos implementados podem ser supervisionados ou sem supervisão, basta que o desenvolvedor programe de acordo com as necessidades do problema. Entretanto, segundo [Mahesh \(2020\)](#), é válido informar que alguns pesquisadores relatam que modelos de aprendizagem de máquina não supervisionados obtêm melhores resultados com uma grande densidade de dados.

O presente trabalho, possui o propósito de envolver a área de tecnologia da informação e a área da saúde. Ao utilizar de métodos e técnicas de aprendizagem máquina, para contribuir na solução de um dos problemas identificados na revisão da literatura, modelos para prognóstico de doenças, mais precisamente, o prognóstico para Septicemia.

Ademais, uma das tecnologias desenvolvidas nos últimos anos, é a que engloba métodos e técnicas que provêm da utilização de modelos estatísticos. Desse modo, diversos estudos já demonstraram que a utilização dessas ferramentas são de grande ajuda no desenvolvimento humano. Trabalho com inteligência artificial e aprendizado de máquina são desenvolvidos, para auxiliar na área da saúde [Kaur et al. \(2020\)](#), não só para sepses como em outras áreas específicas da medicina.

Dessa maneira, têm-se os estudos dos pesquisadores [Mohan et al. \(2019\)](#), que relatam o efeito do uso da predição com aprendizagem de máquina em doenças cardíacas. Além disso, é possível também aplicar em outras áreas como reconhecimento de imagem, auxílio em cirurgias, entre outras soluções que o desenvolvimento tecnológico proporciona, e utilização na produção em larga escala na indústria.

Desse modo, por meio da capacidade de classificação e otimização na separação de outliers ou ruídos, os autores [Widodo and Yang \(2007\)](#) utilizaram da técnica Máquina de Vetores de Suporte (SVM), devido a possibilidade de variações na sua implementação, e relataram a capacidade de diagnóstico na classificação com diversos parâmetros (multiclasses), em uma variedade de áreas de aplicação.

Com isso, pesquisadores no Brasil já aplicam tecnologias de aprendizado de máquina, para prever situações no âmbito hospitalar. Como forma de ilustração, há o estudo sobre a predição do risco de morte em idosos na cidade de São Paulo [mlp \(2020\)](#). Ao utilizar os modelos de aprendizagem de máquina, os pesquisadores ainda puderam analisar a acurácia dos resultados, ao obter uma classificação superior a 0.70 na curva ROC (Característica de Operação do Receptor).

Isto posto, após a submissão e aprovação deste trabalho no âmbito hospitalar, que passará por uma avaliação de especialistas, o melhor modelo desenvolvido estará disponível para ser utilizado em casos reais, em um formato com interface facilitada. Desse modo, como um dos principais frameworks para prognóstico de pacientes está disponível numa versão web [ICU Calculators](#). A disponibilização de uma interface através do modelo de prognóstico criado, se faz altamente necessária.

Assim, para o desenvolvimento do modelo proposto, utilizou-se de tecnologias com bastante

popularidade e suporte pela comunidade de programação. Ao utilizar de linguagens comumente presente em trabalhos de aprendizagem de máquina, como a linguagem Python, e a utilização de ferramentas tecnológicas, como o Google Collaboratory.

Capítulo 4

Relato do Problema

Este capítulo descreve como a identificação do problema de prognóstico de pacientes com septicemia se desenvolveu, e investiga a não aplicação de sistemas existentes para prognosticar a doença em específico.

4.1 Definição

A identificação do problema de prognóstico, se deu a partir da busca na literatura por métodos preditivos no diagnóstico de doenças infecciosas, especificadamente sepses. Após o entendimento do diagnóstico, e também das técnicas utilizadas na predição de diagnóstico, constatou-se métodos utilizados para prognóstico e avaliações de risco de pacientes presentes na UTI, sem especificação na doença do paciente, que são o APACHE e o SAPS.

Mesmo compreendendo que os métodos dependem de uma análise estatística, e contêm o método de regressão logística para obtenção do prognóstico em seu modelo base. Com isso, considerou-se uma hipótese, de que a utilização de ambos os scores, ou de um deles, seria suficiente para elaborar um novo modelo preditivo de prognóstico, com aprendizagem de máquina através da programação.

Não obstante, no âmbito da literatura, também foi identificado a necessidade da diminuição do número de leitos ocupados por pacientes acometidos com septicemia. Além disso, outro fator que potencializa a relevância para a atual pesquisa é a urgência no prognóstico de pacientes com sepses, afinal, quando diagnosticada, há o devido encaminhamento, tratamento, bem como a liberação do paciente conforme a identificação através de um método preditivo.

4.2 SAPS-3

[Moreno et al. \(2005\)](#) desenvolveram um modelo para análise de risco de pacientes, a fim de restringir a generalização dos resultados, com mais de 16,000 pacientes e mais de 300 unidades de tratamento, desde o final de 2002 até 2005. Utilizaram regressão logística para melhorar

o modelo com 20 variáveis, assim foi obtido 0.848 de acurácia da curva ROC, com calibração Hosmer-Lemeshow H= 10.56, P=0.39, C=14.29, P=0.16, e portanto concluíram o modelo preditivo.

Isto posto, a pontuação do método varia de 0 a 217 pontos na elaboração do modelo, comprovando que o mínimo de pontuação atingida foi 5 e o máximo 124. A relação da pontuação SAPS e do estado de vida do paciente se dá pela equação 4.1. Isto é, a probabilidade de morte é resolvida pela equação 4.2.

$$\text{Logit} = -32.6659 + \ln(\text{SAPS3score} + 20.5958) \times 7.3068 \quad (4.1)$$

$$\text{Probabilidadedemorte} = \text{elogit} / (1 + \text{elogit})$$

(4.2)

4.3 APACHE IV

Após a evolução do método APACHE, o APACHE IV foi desenvolvido por [Zimmerman et al. \(2006\)](#), o qual implementou uma regressão logística múltipla, para obtenção da probabilidade de morte nos hospitais. O modelo utilizou-se de 60% da base de dados como treino, e os outros 40% como teste.

Nessa perspectiva, um framework foi desenvolvido para melhor inserção das variáveis do modelo, isto é, 129 variáveis são utilizadas no modelo de regressão logística múltipla do APACHE IV [Ghorbani et al. \(2017\)](#): a idade, sexo, datas de admissão, alta ou óbito, pressão arterial sistólica e diastólica, temperatura corporal, frequência cardíaca, frequência respiratória, glicemia, uréia, soro, creatinina, hematócrito, glóbulos brancos, albumina sérica e bilirrubina, débito urinário durante as primeiras 24 horas de internação na UTI, ph, fração inspirada de oxigênio (FiO2), pressão parcial de dióxido de carbono (PaCO2), pressão parcial de oxigênio (PaO2), bicarbonato (HCO3), escala de coma de Glasgow, ventilação mecânica (VM) e diagnóstico da admissão.

Sendo assim, a acurácia do modelo proposto, através do método do ROC, obteve AUC= 0.88, calibração por (Hosmer- Lemeshow C statistic = 16.9, p = .08). Numa comparação de análise de risco de mortalidade, os autores [Zimmerman et al. \(2006\)](#) observaram que o risco de morte para pessoas com septicemia, através de infecção urinária, atingiu 37.4%, com dados de 2002 à 2003, dentre os 44.288 casos, uma redução de mais de 8% a menos se comparado ao APACHE III, segundo o mesmo autor, que tinha uma taxa de risco de 45.2%.

Nessa perspectiva, esse é o único modelo que tratou especificadamente a análise de risco para pacientes com Septicemia (apesar de avaliar somente um tipo de Sepses) através da infecção urinária.

Por conseguinte, com a utilização do score SAPS de acordo com a tabela 4.1, é observado a variação dos scores obtidos nos experimentos do APACHE IV, Com isso, houve a obtenção da probabilidade de morte através do score SAPS, que variou de 20 até 100, com 95% de intervalo de confiança. Além de ter resultados com a taxa de mortalidade, de acordo com o tempo de internação do paciente, por meio da correlação entre as variáveis, têm-se os dados para aplicar nas hipóteses propostas por este trabalho.

	Taxa de mortalidade %	Varição para baixo %	Varição para cima %
SAPS			
20	2.7	2.0	3.8
30	4.9	3.6	6.6
40	7.8	5.8	10.6
50	11.5	8.5	15.2
60	16.1	12.1	21.0
70	21.8	16.7	28.0
80	28.8	22.4	36.0
90	36.7	29.3	44.8
100	45.5	37.2	54.0
Tempo de Internação			
0	7.1	5.2	9.7
4 hrs	6.5	4.8	8.8
12 hrs	7.2	5.3	9.7
1 day	8.2	6.0	11.0
2 days	9.4	6.9	12.6
4 days	10.5	7.8	14.2

Tabela 4.1: Taxa de mortalidade pelo SAPS e Tempo de internação. Adaptado de [Zimmerman et al. \(2006\)](#)

4.4 Trabalhos Relacionados

Esta seção aborda trabalhos, pelo qual os pesquisadores elaboraram sistemas, a partir da utilização de algoritmos de aprendizagem de máquina. A seção descreve quais as metodologias utilizadas para validação dos modelos desenvolvidos pelos pesquisadores.

4.4.1 Diagnóstico

Segundo o trabalho de [L. et al. \(2021\)](#), propuseram a elaboração de um modelos de diagnóstico, sob a perspectiva de pacientes, que pudessem ser diagnosticado com sepses no menor

intervalo de tempo possível, para poder antecipar o tratamento, e conseqüentemente ajudar na diminuição da mortalidade por sepse.

Desse modo, o estudo foi desenvolvido com técnicas de inteligência artificial, como também de aprendizado de máquina e processamento de linguagem natural. O trabalho analisa a base de dados médicos. Os autores classificaram como dados estruturados, ao tratar de dados comprobatórios que advêm de exames laboratoriais. Como também denominaram como dados desestruturados, informações provenientes da perspectiva dos especialistas, como anotações e prescrições médicas.

Entretanto, ao tratar de dados não estruturados, a avaliação médica de imagens, por parte dos médicos, é tida como uma avaliação bastante precisa. Além disso, existe também a aplicação dos sistemas SIRS e qSOFA, que os auxiliam para um melhor diagnóstico.

No entanto, muitas das vezes os médicos dependem exclusivamente de exames laboratoriais a partir da cultura do sangue do paciente, ou biológicos, que pode demandar tempo o bastante para o agravamento dos sintomas nos pacientes, ainda nesse estudo, a metodologia aplicada para predição é baseada na coleta de sangue, mais precisamente uma gota, no qual obtém-se dados essenciais para o diagnóstico da sepse, principalmente para especificar o tipo da sepse e conseqüentemente ter um melhor prognóstico.

Isto posto, diversos trabalhos levaram em conta dados sociais, histórico de doenças, idade, sexo [Lavoisier et al. \(2016\)](#). Outro trabalho como o de [Ellett et al. \(2018\)](#) que utilizaram dados laboratoriais ao coletar amostras de sangue, principalmente, o paciente que é monitorado na UTI quanto a seu estado atual.

Pesquisadores [van Wyk et al. \(2017\)](#), que priorizaram, além a coleta de dados em um determinado tempo, determinaram outras variáveis que foram utilizadas no modelo de predição, como pressão arterial, pressão arterial média, pressão arterial sistólica, pressão arterial diastólica, frequência cardíaca apical, saturação do oxigênio e fração de oxigênio inspirado.

Sendo assim, diversos pesquisadores levaram em consideração outros sintomas, a exemplo [Mao et al. \(2018\)](#) responsável por analisar a baixa temperatura dos corpos. Com isso, notou-se que pacientes que tiveram choque séptico possuíam pressão de sangue sistólica maior que 90mm Hg, ou seja, pacientes com o quadro de hipertensão depois do diagnóstico de sepse, entraram em estado de choque séptico, isto é, falência múltipla de órgãos, aumentando a criticidade e a probabilidade do paciente não sobreviver.

Por fim, há estudos como o de [Taneja et al. \(2021\)](#), que exploraram as variáveis utilizadas em sua pesquisa, para poder implementar modelos de aprendizagem de máquina, com o intuito de obter resultados quanto ao diagnóstico e prognóstico de pacientes com sepse.

Isto posto, na pesquisa utilizou-se do algoritmo árvores randômicas, para ambos os objetivos propostos pelo trabalho. Com isso, obtiveram resultados com AUC = 0.83 da curva ROC para diagnóstico dos pacientes com septicemia. Já para o modelo de prognóstico, obteve significância com p-value < 0.0001.

4.4.2 Prognóstico

Uma revisão sistemática elaborada pelos [Vellido et al. \(2018\)](#), no qual aborda a relevância da aplicação de algoritmos de aprendizagem de máquina para prognóstico de mortalidade e pacientes. O trabalho descreve os resultados que obtiveram, com aplicação de Regressão logística, RVM (Máquina de vetores de relevância) e SVM com AUC igual a 0.78, 0.86 e 0.89, respectivamente. Com isso, a revisão utilizou-se de métodos comparativos da literatura, como SIRS e APACHE-II.

[Hu et al. \(2022\)](#) analisaram a eficiência da predição de mortalidade de pacientes, através de mais de 8 mil dados de pacientes acometidos com septicemia. Os autores utilizaram um método de exclusão de dados, no qual foi responsável por remover mais de 800 pacientes que não permaneceram internados por pelo menos 24h.

Por fim, os autores conseguiram realizar a implementação dos algoritmos de aprendizagem de máquina *eXtreme Gradient Boosting* (XGBoost), Árvore de decisão, *Naive Bayes* (NB), Floresta Randômica (RF) e regressão logística (LR). Assim, concluíram a relevância do trabalho com o melhor resultado para o modelo XGBoost, com AUC= 0.884.

Os pesquisadores [Silva Junior et al. \(2010\)](#), utilizaram o método SAPS-3 ¹ em hospitais Brasileiros, e comprovaram a eficiência do método ao aplicar em 16.784 pacientes presentes em mais de 300 UTI's. Com isso, ao analisar um bom poder discriminatório dos pacientes, que sobreviveram dos que não sobreviveram, os autores perceberam que o sistema SAPS não foi desenvolvido para áreas específicas, uma vez que eles utilizaram o sistema de maneira geral e salientaram a importância da validação em pacientes com diagnósticos específicos.

Com a aplicação do modelo, puderam analisar um total de 1310 pacientes, o método SAPS-3 com pontuação 57 obteve melhores resultados com 75% de sensibilidade, 86% de especificidade e com AUC = 0.86. Os pesquisadores relataram que houve uma facilidade de implementação do modelo SAPS-3 no hospital, devido a simplicidade de calcular as variáveis, além de compará-los com os métodos anteriores, como o APACHE-2 e SAPS-2.

Ademais, no Brasil, a utilização do framework APACHE IV apresentou resultados satisfatórios, como o trabalho de [de Carvalho et al. \(2020\)](#), que teve como objetivo utilizar a calculadora online do APACHE IV, disponível no site [Framework para cálculo do APACHE IV \(2022\)](#). Sendo assim, ao alimentar o framework com os dados disponibilizados pelo hospital que atuaram, com ROC ≥ 0.80 , foi observado também com uma proposta semelhante a deste trabalho, isto é, uma análise do tempo de permanência dos pacientes na UTI através do coeficiente de Spearman, no qual foi detectada uma correlação fraca entre o score APACHE IV e o tempo de internação ($r = 0,025$; $p < 0,0001$).

Por outro lado, os pesquisadores [Zhai et al. \(2020\)](#) conseguiram realizar uma comparação entre os sistemas criados a partir dos algoritmos de aprendizagem de máquina, ao comparar os modelos para predição de mortalidade em pacientes na UTI, mais precisamente com casos

¹ Simplified Acute Physiology Score [Moreno et al. \(2005\)](#)

mais críticos.

Os pesquisadores compararam os modelos Máquina de Vetores de Suporte (SVM), GBDT, XGBoost e Regressão Logística, os modelos obtiveram os resultados de acurácia pela Área Sob a Curva (AUC) de 0.794, 0.840, 0.849 e 0.822, respectivamente, em comparação com o método SAPS-3 com 0.826 de AUC.

Por fim, com uma melhor elucidação quanto as metodologias utilizadas em cada trabalho, e como forma de obter uma análise quantitativa, está presente na tabela 4.2 as metodologias de cada trabalho em comparação ao presente trabalho.

Trabalhos	Parâmetros	Melhor Algoritmo	Melhor AUC
Vellido et al. (2018)	(SOFA e SAPS na UTI admissão, bem como seus valores mínimos e máximos correspondentes durante a UTI fique).	SVM	0.89
Hu et al. (2022)	Característica Demográfica; Hábitos/Histórico familiar; Sinais Vitais; achados laboratoriais; Tratamento Médico; Função Neurológica (Escala de Coma Gasglow);	XGBoost	0.88
Silva Junior et al. (2010)	Variáveis aplicadas no SAPS-3	SAPS-3	0.86
de Carvalho et al. (2020)	Variáveis do APACHE-IV	APACHE-IV	0.80
Zhai et al. (2020)	dados demográficos dos pacientes extraídos; dados fisiológicos; testes laboratoriais; diagnóstico; tempo de permanência; comorbidades;	XGBoost	0.849
Presente Trabalho Bento M.V.S	SAPS-3 prognóstico; gênero; estado de vida do paciente; e permanência do paciente na UTI (>48h e <48h);	SVC	0.85

Tabela 4.2: Comparação dos algoritmos dos trabalhos quanto a prognóstico de Mortalidade.

Capítulo 5

Proposta

Neste capítulo está descrito a proposta do trabalho, bem como os detalhes a serem considerados na elaboração do modelo de aprendizagem de máquina, e as correlações das variáveis utilizadas.

5.1 Fundamentação

A fim de desenvolver um modelo a partir de algoritmos de aprendizado de máquina. Em primeiro lugar, buscou-se na literatura, trabalhos relacionados, que disponibilizassem informações que facilitassem a implementação dos modelos de aprendizado de máquina.

Nessa perspectiva, os modelos foram desenvolvidos a partir de dados de especialistas intensivistas de telemedicina que fornecem suporte médico a médico de forma remota, o que, conseqüentemente, contribui para o prognóstico da doença.

Isto posto, tem-se como objetivo técnico a utilização e comparação de modelos de aprendizagem de máquina, a fim de obter o melhor modelo para predição no prognóstico dos pacientes, com resposta para o risco de mortalidade e tempo de internação do paciente.

Inspirado no estudo de caso [Balkan et al. \(2018\)](#), os modelos que foram elaborados por este trabalho têm como fim a elaboração de um framework, com a possibilidade de obter uma entrada de dados, no qual busca uma melhora do prognóstico de pacientes acometidos com septicemia.

Este trabalho utilizou-se dos modelos da literatura SAPS-3 e APACHE IV, como referência na comparação quantitativa. Pela falta de uma validação dos Entretanto, pela falta de uma validação dos especialistas, principalmente por não possuírem um framework com interface gráfica, não foi possível disponibilizar um modelo como uma API (Interface de Programação de Aplicação), assim como o framework APACHE IV utilizado neste endereço [Framework para cálculo do APACHE IV \(2022\)](#).

O framework APACHE IV, é utilizado para avaliar os riscos do paciente, ao dar prognóstico de mortalidade de pacientes, pontuação APS, referente ao SAPS-3, a taxa de probabilidade de

morte e a taxa de probabilidade de tempo de internação, ver figura 5.1.

APACHE IV Score	<input type="text"/>	/286
APS Score	<input type="text"/>	/239
Estimated Mortality Rate	<input type="text"/>	%
Estimated Length of Stay	<input type="text"/>	days

Figura 5.1: Resultado da calculadora web APACHE IV ICU Calculators Framework para cálculo do APACHE IV (2022)

Os dados utilizados na elaboração dos modelos, são provenientes de pacientes que foram admitidos em mais de 400 hospitais, segundo os dados clínicos presentes no dataset disponível em [eICU Collaborative Research Database \(2022\)](#). Com dados reais e relevantes que são primordiais para a elaboração do modelo preditivo proposto por esse trabalho, além dos dados gerais de pacientes, há também a disponibilização de resultados da metodologia Apache IV, e APS, como foi citado anteriormente, dados que serão comprobatórios e que servem de validação.

Sendo assim, durante o desenvolvimento do modelo, primeiramente foi observado e definido os dados clínicos que faz sentido para a elaboração do modelo, estabelecer uma correlação adequada para os modelos de aprendizagem de máquina. Nessa perspectiva, é importante citar que nessas informações os pacientes são identificados com uma identidade única, assim é possível observar a evolução do paciente, informação fundamento para elaboração de um modelo de prognóstico.

Desse modo, foi disponibilizado pela base de dados as informações admissionais do paciente, como o diagnóstico inicial. Foi observado qual o diagnóstico posterior ao período de 24h que o paciente foi admitido. Por fim, obter as pontuações do apache IV e SAPS, como também avaliar o estado do paciente de vivo ou expirado (atribui-se aos pacientes que vieram a óbito).

5.1.1 Fluxo de normalização dos dados

O fluxo desenvolvido ao longo do desenvolvimento do presente trabalho, é demonstrado na figura 5.2, para ser aplicado especificadamente em pacientes que foram diagnosticados, ou, que desenvolveram septicemia ao longo de sua internação hospitalar.

Foi necessário a separação dos dados, referente a base de informações disponibilizadas pelos pesquisadores [Goldberger et al. \(2000\)](#) e [Johnson et al. \(2021\)](#), conforme a identificação de pacientes duplicados, ou outliers desnecessários para a implementação dos modelos propostos.

Portanto, a normalização foi efetuada de acordo com as tabelas disponíveis na base de dados, e a correlação entre as tabelas se deu pelas variáveis únicas de identificação do paci-

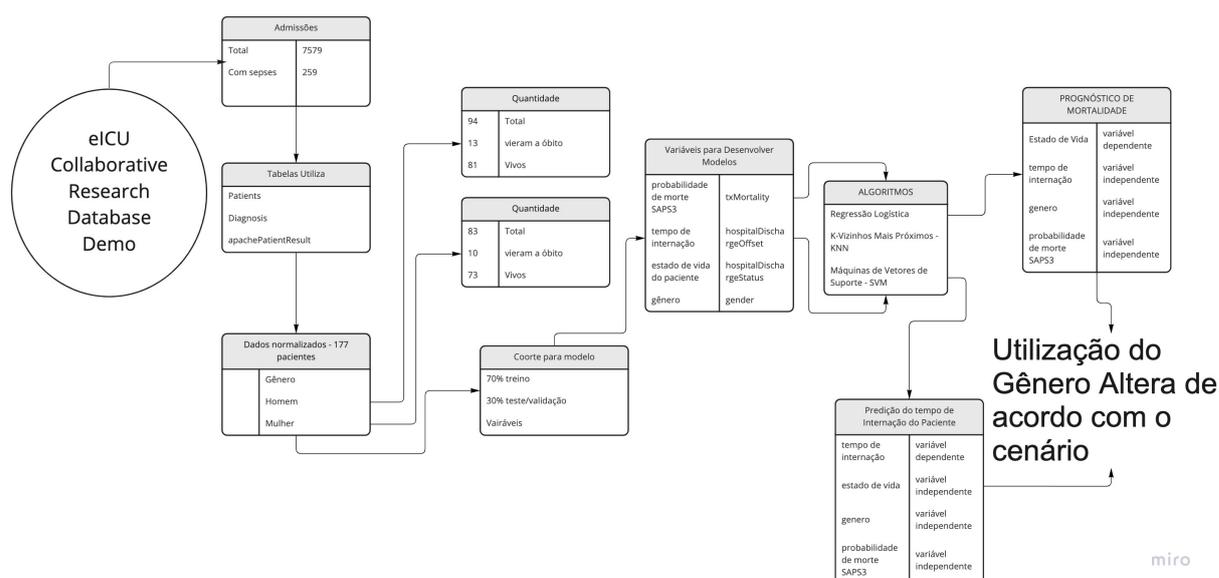


Figura 5.2: Fluxo aplicado no desenvolvimento dos modelos

ente, bem como o seu diagnóstico, estado de vida atual, tempo de internação e pontuação dos métodos APACHE-IV e APS.

5.2 Hipóteses

Esta seção reúne e descreve as hipóteses que foram elaboradas no decorrer da revisão da literatura, desde os modelos que serviram para diagnóstico, até os modelos para prognóstico.

As hipóteses que foram elaboradas a partir da revisão sistemática, a resolução final e as que foram testadas, foram as hipóteses abaixo listadas.

- É possível obter prognóstico eficaz quanto a pacientes com septicemia.
- Aplicação de algoritmos de aprendizagem de máquina, obtém resultados iguais ou superiores de AUC, quanto aos métodos APACHE- IV e SAPS-3.
- Aplicação algoritmos de aprendizagem de máquina, predizem probabilidade de óbito de pacientes com septicemia.
- Aplicação algoritmos de aprendizagem de máquina, predizem probabilidade do tempo de internação do paciente na UTI.

Nessa perspectiva, ambas as hipóteses possuem características semelhantes quanto ao tempo de coleta de dados para obtenção de seus scores, logo, foi priorizados pacientes que possuíam tempo de internação superior a 24h, de acordo com uma das variáveis presentes da base de dados.

Isto posto, implementado 3 algoritmos de aprendizagem de máquina, ao levar em consideração a base de dados obtida. Com isso, os algoritmos implementados foram: KNN, SVM e o algoritmo de Regressão logística. A justificativa para aplicação dos três algoritmos, é baseada na necessidade de classificar o paciente, para isso, o trabalho teve que usar modelos classificatórios e probabilísticos. Ambos os algoritmos foram implementados com 70% treino e 30% teste da base de dados obtida.

Além disso foi calculada a pontuação do SAPS-3, para utilização como variável independente dos modelos, a partir do score disponibilizado na base de dados e obter a probabilidade de morte como forte correlação das variáveis.

Portanto, para atingir os objetivos principais do presente trabalho, as hipóteses foram elaboradas para construção de um modelo preditivo para prognóstico, a fim de diminuir o tempo de execução por parte dos especialistas quanto as tomadas de decisões referente a liberação de leitos das UTI's e aumentar a taxa de sucesso da sobrevivência do paciente. Não somente isso como também há o desejo de contribuir na diminuição do tempo de internação.

Capítulo 6

Metodologia de Experimento

Este capítulo descreve a metodologia que será aplicada no sistema, como será desenvolvido, técnicas, instrumentos, ou, dispositivos que serão utilizados para obtenção dos resultados, ao desenvolver um modelo de prognóstico para pacientes com sepses.

6.1 Métodos

Na busca de utilizar dados reais, com intuito de validar os modelos implementados, foi encontrado um dataset elaborado por [Goldberger et al. \(2000\)](#) e [Johnson et al. \(2021\)](#), Portanto, os números fazem parte de um grupo de pesquisadores especialistas em telemedicina [Pollard et al. \(2018\)](#), que contabilizaram mais de 200 mil admissões monitoradas nos estados dos Estados Unidos entre os anos 2014 e 2015. A base de dados foi publicada em 2021 com as nomenclaturas estabelecidas pela [Brooks and Tsao](#) e encontrado no site [eICU Collaborative Research Database \(2022\)](#), de maneira *open-access*.

Nesse interím, os dados dos pacientes são provenientes de mais de 2500 unidades de internações e de aproximadamente 20 hospitais, isto é, as características dos pacientes estão descritas na tabela 7.3. Desse modo, as tabelas analisadas para este trabalho estão descritas na tabela 6.1, e foi possível obter a normalização dos dados obtidos, principalmente pela disponibilização de uma identificação única dos pacientes. É válido informa, que os dados disponíveis não dispõem de informações pessoais ou privadas dos pacientes.

As tabelas analisadas para este trabalho estão descritas na tabela 7.1, assim foi possível a normalização dos dados obtidos, principalmente pela disponibilização de uma identificação única dos pacientes.

No que diz respeito à parte técnica da programação, foi utilizada a linguagem Python no desenvolvimento dos modelos, com a disponibilidade de algumas ferramentas para Python, isto é, os algoritmos foram implementados pela ferramenta do Google, o Google collab. Desse modo, os algoritmos foram implementados com mais facilidade devido à importação da biblioteca do *scikitLearn* [Pedregosa et al. \(2011\)](#) de KNN, SVM e Regressão Logística.

Estratégia

A estratégia desenvolvida consiste em:

- Separar os pacientes de todos os 20 hospitais.
- Separar os pacientes que foram admitidos com sepse para modelo de treino.
- Separar os pacientes que foram diagnosticados com sepse/doença infecciosa com variável major sendo priority ou primary.
- Separar os pacientes com a variável da coluna apachePatientResult - segundo a coluna da base de dados-, que foram diagnosticados com sepsis.
- Coleta de Dados clínicos de UTI's.
- Elaboração de um sistema com modelo de aprendizagem de máquina.
- Fazer análise quantitativa dos modelos de aprendizagem de máquina para predição.
- Predizer o prognóstico de mortalidade a partir do modelo criado.
- Contribuir para a diminuição da mortalidade por sepses.

Ademais, com a estratégia para a elaboração do modelo, deve-se diferenciar as admissões dos diagnósticos - se houver - para poder validar o modelo de treino. Este trabalho também segue a metodologia dos artigos citados, tais como [Balkan et al. \(2018\)](#) ao utilizar entre 67% - 70% para treino e entre 30%-33% para teste no modelo elaborado. Portanto, a presente pesquisa, utilizou 30% para teste, e 70% para treino, da base de dados obtida.

Capítulo 7

Validação

Este capítulo descreve os meios de validação dos modelos de aprendizagem de máquina elaborados por este trabalho, e justifica a não validação com os especialistas, além de não ter sido possível a aplicação dos modelos *in-loco* nos hospitais.

7.1 Ferramentas

Para elaboração dos modelos de aprendizagem de máquina, utilizou-se do Google collaboratory, com a utilização da linguagem Python. A utilização dessa ferramenta, é facilitada através da capacidade de processamento de dados, além de outros processos computacionais, que são de responsabilidade da plataforma.

Mesmo que haja um processamento por parte da plataforma, vale ressaltar que toda a elaboração e os testes serão realizados através de um Macbook Pro Early 2015, com processador 2,7 GHz Dual-Core Intel Core i5, memória RAM de 8 GB 1867 MHz DDR3, Intel Iris Graphics 6100 1536 MB e armazenamento SSD -SAMSUNG EVO 980- de 500GB.

7.2 Análise dos resultados

Esta seção descreve os métodos que foram utilizados, para avaliar os resultados dos modelos implementados, e realizar uma comparação quantitativa com os modelos da literatura, de acordo com as metodologias por eles utilizadas, com os modelos estatísticos.

7.2.1 Característica de Operação do Receptor - curva ROC

Utilizou-se a curva ROC como modelo estatístico, para avaliar quantitativamente os modelos propostos por este trabalho, pois os modelos presentes na literatura, APACHE e SAPS, apresentaram esse modelo como forma de validação dos seus respectivos métodos.

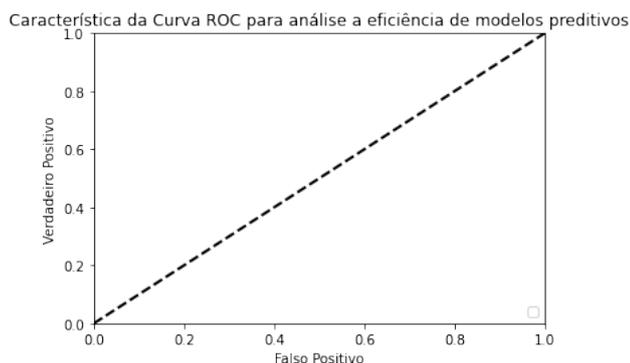


Figura 7.1: Exemplo Curva ROC

Um exemplo do método está presente na figura 7.1, desenvolvido enquanto se implementava os modelos de aprendizagem de máquina. A curva ROC [Bradley \(1997\)](#), é um validador, pelo qual os algoritmos de aprendizagem de máquina são mensurados a partir da sua acurácia. E determina assim, a eficiência do modelo quanto a sua sensibilidade e especificidade para o problema proposto, ou, verdadeiros positivos e falsos positivos.

7.3 Cenário

Esta seção aborda os seguintes cenários: cenários 1 e 2, para predição de mortalidade, com base em variáveis diferentes; e cenário 3 para predição de tempo de internação para menos de 48h e mais de 48h.

Com relação à obtenção do melhor desempenho e aplicação do modelo de aprendizagem de máquina, o cenário ideal para este trabalho é o ambiente hospitalar, mais especificamente o setor da UTI, no qual são obtidos dados significativos que irão contribuir para a elaboração do modelo de teste, como também dados dos critérios previamente relatados, SAPS e APACHE.

Sendo assim, o cenário ideal é necessário, uma vez que por meio da simulação não é possível ter uma relevância na aplicação deste trabalho porque um dos objetivos é a aplicação em ambientes hospitalares, para então auxiliar médicos e enfermeiros com suas metodologias e critérios de avaliação.

Desse modo, utilizou-se de técnicas para que os dados disponíveis favorecessem ao máximo os objetivos deste trabalho. Assim, o cenário contém os pacientes que foram diagnosticados após a admissão, pois na própria base há pessoas admitidas com sepse, além disso, o cenário foi nor-

malizado para poder conter as variáveis utilizadas nos modelos de aprendizagem de máquina, também é devido a pontuação SAPS ser mais eficiente após 24h de tempo de internação, assim pode-se avaliar melhor os resultados de prognóstico.

Tabela	Descrição
admissionDx	Contém o diagnóstico primário para admissão na UTI de acordo com os critérios de pontuação APACHE. Inserido nos formulários de notas do paciente. Após um período fixo de admissão, a tabela não pode ser atualizada pelo cuidador.
diagnosis	Diagnóstico do paciente registrado na tabela de diagnóstico ativo. A sequência indica gravidade relativa. A prioridade de diagnóstico não é necessária.
patients	Contém dados demográficos do paciente e detalhes de admissão e alta para estadias no hospital e na UTI.
apachePatientResult	Fornece previsões feitas pelo escore APACHE (versões IV e IVa), incluindo probabilidade de mortalidade, tempo de internação e dias de ventilação.
treatment	Permite documentar, de forma estruturada, tratamentos ativos específicos para o paciente.

Tabela 7.1: Tabelas Disponíveis com dados clínicos dos pacientes.

Base de dados

A disponibilização dos dados é feita através de arquivos comprimidos com tabelas que contêm os tipos de dados clínicos que referenciam os pacientes em Comma-separated values (CSV), ou arquivos individuais das tabelas em CSV, assim é possível selecionar as tabelas que serão relevantes para a pesquisa.

Portanto, foram utilizadas algumas das 31 tabelas disponibilizadas por essa base de dados, para separar as variáveis conforme o modelo de aprendizagem de máquina necessita, conforme a tabela 6.1, que descreve cada arquivo utilizado para este trabalho de forma adaptada conforme as definições presentes no site [eICU Collaborative Research Database \(2022\)](#)

A disponibilização dos dados é feita através de arquivos comprimidos com tabelas que contêm os tipos de dados clínicos que referenciam os pacientes em Comma-separated values (CSV), ou arquivos individuais das tabelas em CSV, assim é possível selecionar as tabelas que serão relevantes para a pesquisa.

Portanto, foram utilizadas algumas das 31 tabelas disponibilizadas por essa base de dados, para separar as variáveis conforme o modelo de aprendizagem de máquina necessita, conforme a tabela 7.1, que descreve cada arquivo utilizado para este trabalho de forma adaptada conforme as definições presentes no site [eICU Collaborative Research Database \(2022\)](#)

Os dados clínicos que foram utilizados de cada arquivo, está presente na tabela 7.1, na qual consta as colunas com valores das variáveis extraídas para elaboração do modelo de aprendizagem de máquina. A variável `patientunitstayid`, é a coluna em comum em todos os arquivos, pois é uma variável que contém as identificações dos pacientes, com ela é possível saber todos os dados clínicos, como por exemplo gênero e idade de determinada pessoa.

Coluna	Descrição Coluna
admissionDx	patientunitstayid, admitdxttext e admitdxttext
diagnosis	patientunitstayid, activeupon-discharge, diagnosisstring e apacheadmissiondx
patients	patientunitstayid, gender, apacheadmissiondx, hospital-dischargestatus, unitdischargestatus e hospitalDischargeStatus
apachePatientResult	patientunitstayid e acutephysiologyscore,
hospitalDischargeOffset	Tempo de internação do paciente em minutos.
hospitalDischargeStatus	Estado de vida do paciente ALIVE = VIVO; EXPIRED= MORTO.

Tabela 7.2: Variáveis da base de dados do eiCU [Brooks and Tsao](#)

Gênero	Quantidade com septicemia	Vivos	Mortos
Mulheres	83	73	10
Homens	94	81	13
Total	177	154	23

Tabela 7.3: Característica dos pacientes internados presente na base de dados.

7.4 Metodologia

Os algoritmos de classificação são em sua maioria utilizados para poder avaliar problemas binariamente, ou seja, há a mesma lógica de classificação instaurada pelos hospitais na utilização dos critérios previamente mencionados.

Nessa perspectiva, ao levar em consideração a base de dados obtida, foi levantada a hipótese da aplicação de 2 algoritmos de aprendizagem de máquina: KNN e Regressão logística. A justificativa para a aplicação dos dois algoritmos é baseada em que precisamos ter modelos de comprovação que sejam classificatórios e probabilísticos. Assim, ambos os algoritmos terão como 70% treino e 30% teste, de acordo com a base de dados. Além disso, será calculada a pontuação do SAPS-3, para utilização como variável independente dos modelos, a partir do score disponibilizado na base de dados.

Desse modo, espera-se prever as chances de um paciente com septicemia morrer, ao considerar o tempo de internação do paciente, bem como a utilização da pontuação de mortalidade do SAPS, do mesmo modo na classificação.

Embora alguns pesquisadores relatem que modelos de aprendizagem de máquina não supervisionados obtenham melhores resultados com uma grande densidade de dados [Mahesh](#)

(2020), há também modelos supervisionados que obtêm resultados com boa validação. Inclusive, o presente trabalho se utiliza de uma implementação dos algoritmos de maneira supervisionada.

7.4.1 KNN - K-Nearest Neighbor

O *KNN* é um algoritmo de classificação não paramétrica. É conhecido pela sua simplicidade e eficácia. Por ser supervisionado necessita de um conjunto de dados de treinamento rotulado, os quais são categorizados em várias classes, de modo que a classe dos dados não rotulados seja prevista Taunk et al. (2019).

Funcionamento

O que queremos fazer com este algoritmo é uma classificação, logo o funcionamento dele se dá pela aproximação do parâmetro K , que é representado pela quantidade de vizinhos de aproximação do ponto referente ao que queremos classificar.

Portanto, iremos classificar de acordo com o número de K 's perto do ponto que queremos a depender do cenário, por exemplo se tivermos $K=3$, o ponto que queremos classificar irá observar quais os 3 vizinhos mais próximos, e a partir disso ter sua classe definida.

7.4.2 RL - Regressão Logística

Um dos modelos estatísticos mais utilizados na literatura, a regressão logística consegue atingir várias áreas como sociais, médicas, econômicas entre outras, das quais se beneficiam pelo seu modelo probabilístico, principalmente por obter resultados com variáveis não correlacionadas.

Um grande fator de impacto é na área médica, ao contribuir com os critérios de pontuação, vide APACHE e SAPS-3, que têm em sua base a regressão logística implementada no seu modelo final.

Regressão logística é um algoritmo preditor, com principal resultados que variam entre 0(zero) à 1. Há a regressão logística com variável independente e independente, como também há a regressão logística múltipla, para os casos em que é necessário o tratamento de múltiplas variáveis, muitas das vezes por não serem correlacionadas.

Funcionamento

A base do funcionamento deste algoritmo, é estimar a probabilidade associada a ocorrência dado um evento. A probabilidade é baseada em um conjunto de variáveis explanatórias, seja de natureza dicotômica(partição de um todo) ou binária. Assim os resultados probabilísticos são definidos no intervalo de 0(zero) à 1.

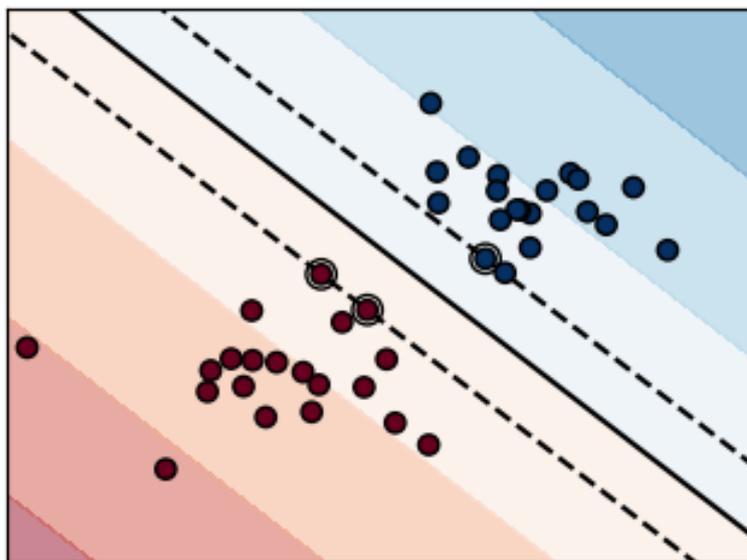


Figura 7.2: Exemplo de Hiperplano - [scikit-learn \(2022\)](#)

7.4.3 SVM - Máquina de Vetores de Suporte

Assim como os modelos anteriores, o SVM é um modelo bastante utilizado pelos cientistas de dados. Proposto pelos autores [Boser et al. \(1992\)](#), o SVM é um algoritmo aplicado a ambientes supervisionados. A base de seu funcionamento está na implementação da função Kernel, onde é responsável pela classificação dos conjuntos de dados, de acordo com o mapeamento dos pontos num espaço multidimensional.

Funcionamento

Com a função Kernel, o SVM atua na separação dos dados com diferentes classes. Assim, o modelo aplica o conceito de planos de decisão, para traçar uma separação de forma linear, conforme as três retas que são postas no hiperplano. É possível observar um exemplo presente na biblioteca [Pedregosa et al. \(2011\)](#), pela figura 7.2.

Na figura é possível visualizar a separação dos conjuntos. As duas linhas mais próximas do conjunto delimitam o hiperplano, e a linha do meio representa a separação dos conjuntos de forma linear. Portanto, com os conjuntos separados e com inserção de novos objetos no plano, será possível classificar os novos objetos de acordo com as delimitações das retas.

Variáveis	Descrição
gênero	gênero do paciente, Homem ou Mulher
diagnóstico	Contém o diagnóstico primário para admissão na UTI de acordo com os critérios de pontuação APACHE. Inserido nos formulários de notas do paciente. Após um período fixo de admissão, a tabela não pode ser atualizada pelo cuidador.
tempo de internação	Diagnóstico do paciente registrado na tabela de diagnóstico ativo. A sequência indica gravidade relativa. A prioridade de diagnóstico não é necessária.
estado do paciente	Contém dados demográficos do paciente e detalhes de admissão e alta para estadias no hospital e na UTI.
probabilidade de morte SAPS-3	Contém dados demográficos do paciente e detalhes de admissão e alta para estadias no hospital e na UTI.

Tabela 7.4: Parâmetros utilizados nos modelos após normalização dos dados.

7.5 Parâmetros

Esta seção descreve os parâmetros que foram utilizados para a elaboração dos modelos de aprendizagem de máquina a partir das hipóteses iniciais na busca por resultados que satisfaçam o problema de pesquisa.

Os parâmetros utilizados por esta pesquisa, são : a probabilidade do paciente vir a óbito morte, estado do paciente, gênero e o tempo de internação. Assim, após passar pela normalização dos dados, e redefinidos, a variável gênero, foi redefinida de Homem e Mulher, para binário, 1 para Homem e 0 para Mulher.

A variável tempo de internação, também foi redefinida pra binária também, com 1 identificando pacientes que permaneceram na UTI maior que 48h, e 0 para tempo de internação menor ou igual a 48h. Como também a variável estado de vida do paciente, foi normalizada de modo binário, com valor 1 para pacientes vivos, e 0 para pacientes que faleceram.

Os parâmetros utilizados como variáveis independentes na elaboração dos modelos para predição de prognóstico de mortalidade, foram as variáveis: gênero, tempo de internação e probabilidade de morte calculada pelo SAPS-3. E para predição de tempo de internação do paciente, foi considerada as variáveis: gênero, estado de vida do paciente e probabilidade de morte calculada pelo SAPS-3.

7.6 Experimentos

Após normalização do dataset dos pesquisadores [Goldberger et al. \(2000\)](#) e [Johnson et al. \(2021\)](#), foi possível obter 177 pacientes diagnosticados com septicemia, dentre eles 83 mulheres

e 94 homens. A estratégia para aplicação dos algoritmos no dataset, foi de utilizar 70% para treino e 30% da base para teste.

Sendo assim, a hipótese gerada para implementação dos modelos de aprendizagem de máquina, consiste em utilizar as equações 4.1 e 4.2, e gerar uma nova coluna que chamamos de txMortality, onde está presente o cálculo do percentual de mortalidade do paciente diagnosticado com sepses pelo SAPS-3, com intervalo de 0(zero) à 1. Foi possível também notar, que os valores encontrados quanto ao percentual da probabilidade de morte pelo SAPS-3, estiveram de acordo com a tabela 3.1 do APACHE-IV, ao realizar testes com entradas hipotéticas das variáveis independentes.

Foi possível também notar, que os valores encontrados quanto ao percentual da probabilidade de morte pelo SAPS-3, estiveram de acordo com a tabela 4.1 do APACHE-IV, ao realizarmos testes com entradas hipotéticas das variáveis independentes.

Capítulo 8

Resultados

Nesta seção descreve os resultados perante aos cenários propostos, bem como a comparação dos dos modelos presentes na literatura, APACHE IV e SAPS-3, com os modelos de aprendizagem de máquina implementados.

Isto posto, os resultados obtidos foram satisfatórios, mesmo que a normalização dos dados tenha afetado significativamente a quantidade de pacientes, ao identificar duplicações. Com isso, ainda sim foi possível aplicar os algoritmos propostos, estes que serão descritos abaixo e foram disponibilizados em : [Marcos Vinícius \(2022b\)](#) e [Marcos Vinícius \(2022c\)](#)

Por conseguinte, os resultados obtidos foram satisfatórios, mesmo que a normalização dos dados tenha afetado a quantidade de pacientes, ainda assim foi possível aplicar os algoritmos propostos, estes, descritos abaixo e disponibilizados em:

- [Marcos Vinícius \(2022a\)](#) *Google Colaboratory*
- [Marcos Vinícius \(2022b\)](#) *GitHub operSource Code*
- [Marcos Vinícius \(2022c\)](#) Publicado na Code Ocean para o *Instituto de Engenheiros Eletricistas e Eletrônicos ou Instituto de Engenheiros Eletrotécnicos e Eletrônicos ou IEEE*

8.1 Cenários

Com os dados normalizados e disponíveis para a implementação dos algoritmos, foi pensado juntamente com as hipóteses, cenários que pudessem atenderem as propostas iniciais, como a utilização das variáveis obrigatórias mortalidade e tempo de internação para atingir o objetivo deste trabalho.

Por conseguinte, a estrutura dos cenários está composta pelos cenários 1 e 2, em que foram implementados os algoritmos para obter resultados preditivos quanto a probabilidade do paciente com sepse estar vivo ou morto. Já nos cenários 3 e 4, foram implementados os mesmos algoritmos com objetivo de obter resultados quanto à estimativa de tempo de internação do paciente na UTI.

8.1.1 Cenário 1

No cenário 1, observamos os algoritmos com os parâmetros tempo de internação, gênero, probabilidade de morte pelo saps, como variáveis independentes, e a o estado de vida do paciente no hospital, como variável dependente.

8.1.2 Regressão Logística

Este modelo, obteve resultados significativos com $AUC = 0.82$, ver figura 8.4, conforme figura (roc curve comparasion) e uma matriz de confusão como pode ser analisada na figura 8.1 a quantidade de pacientes que o modelo conseguiu prever, com 46 predições para pacientes com estado vivo, com 4 não classificados, e com estado morto 2 pacientes preditos e 2 não classificados.

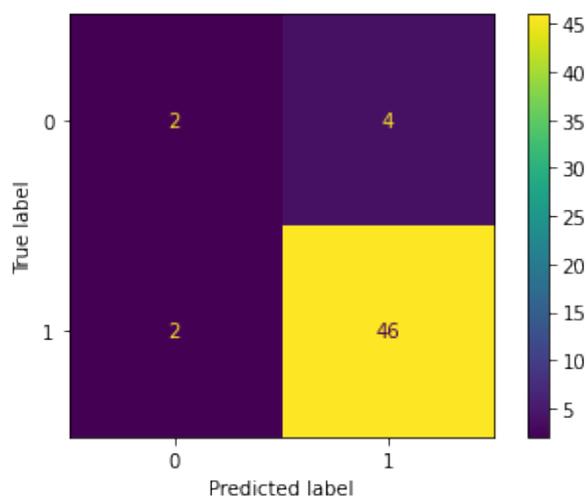


Figura 8.1: Matriz de Confusão Regressão Logística - cenário 1

8.1.3 KNN

Com uma $AUC = 0.74$ de acordo com a curva ROC 8.4, a acurácia com esse modelo, foi a pior dentre as implementadas para este cenário. Também foi gerada uma matriz de confusão presente na figura 8.2, com predição para pacientes vivos igual ao modelo de regressão logística, 46, com 3 pacientes não classificados para o mesmo estado, como também, 2 pacientes com estado de vida falecido, com 3 não classificados.

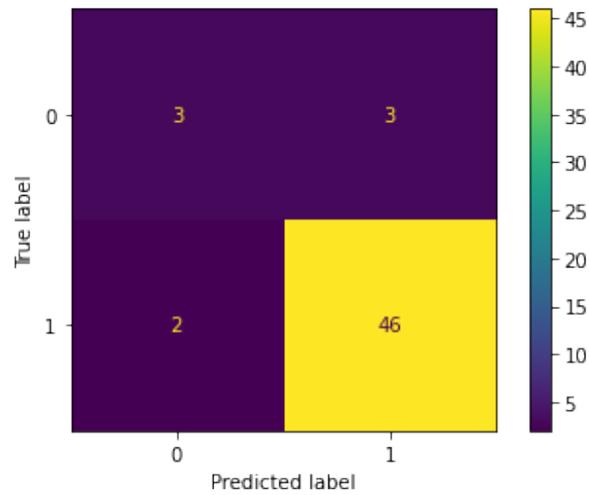


Figura 8.2: Matriz de Confusão KNN - cenário 1

8.1.4 SVM

O SVM obteve maior AUC = 0.85 dentre os modelos implementados, de acordo com a comparação entre as curvas roc 8.4. O modelo possui também uma matriz de confusão semelhante aos algoritmos anteriores, com 46 pacientes com predição correta e 4 não classificados para o estado do paciente vivo, como também 2 com predição correta e 2 não classificados para o estado de vida do paciente falecido.

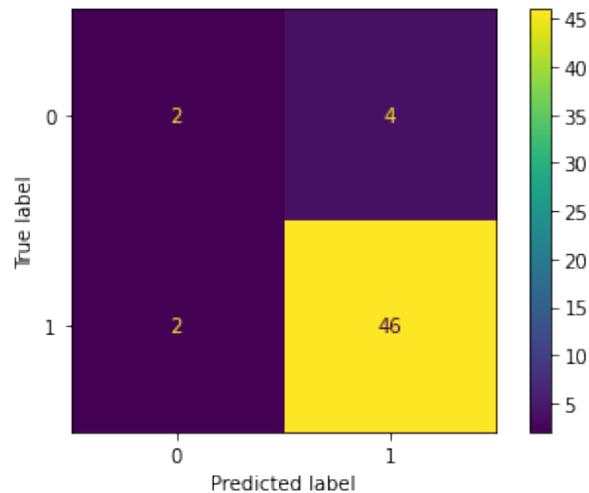


Figura 8.3: Matriz de Confusão SVM - cenário 1

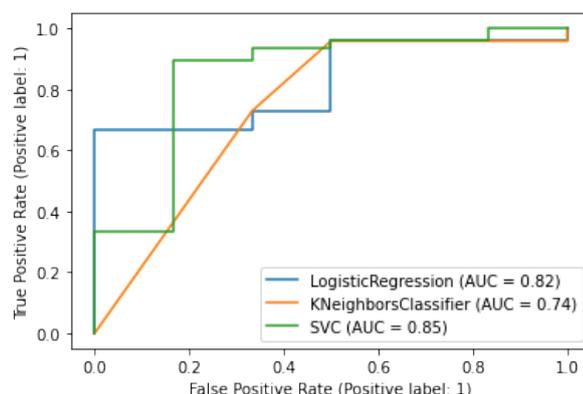


Figura 8.4: Curva ROC dos modelos - cenário 1

8.2 Cenário 2

Já no cenário 2, foi desconsiderado a variável gênero, a fim de analisar os resultados de cada modelo, com isso, foi possível implementar nos algoritmos propostos e obter uma comparação significativa de cada modelo.

Isto posto, este cenário não obteve resultados superiores em nenhum dos modelos, ao ser comparado com o cenário 1, justamente por não se utilizar da variável gênero.

8.2.1 Regressão Logística

Ao avaliar a matriz de confusão do modelo, presente na figura 8.5, percebe-se que a predição correta para estado do paciente vivo, está com uma classificação maior (43), e 8 incorretas, por outro lado, a predição para o estado de morte obteve 2 predições corretas, e 1 não classificada.

Para valiar o grau de sensibilidade e especificidade do modelo, obteve uma AUC = 0.70 ver figura 8.6 que compara com os outros modelos.

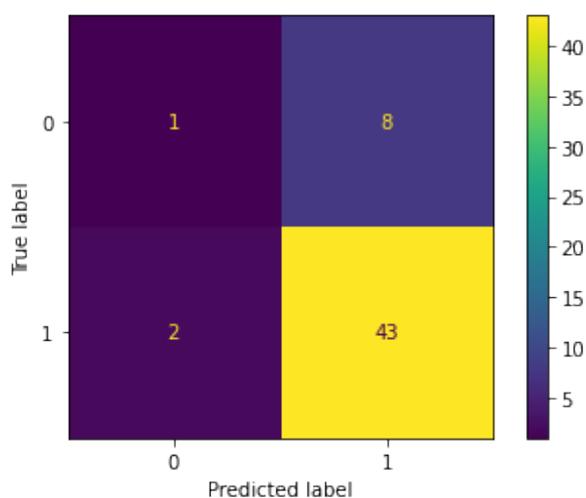


Figura 8.5: Matriz de Confusão Regressão Logística - cenário 2

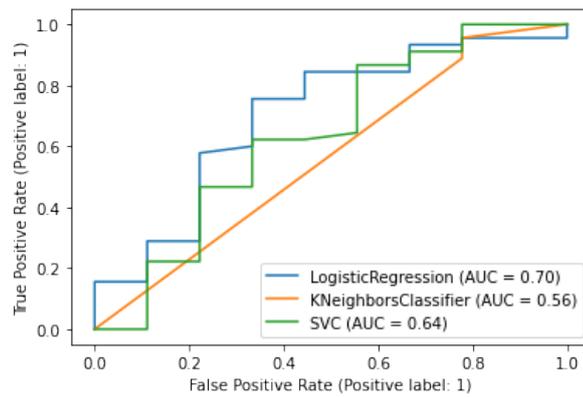


Figura 8.6: Curva ROC dos modelos - cenário 2

8.2.2 KNN

No KNN não foram obtidos resultados significativos ao inferir a variável gênero, como em regressão logística. Assim os resultados obtidos quanto a análise da curva ROC 8.6, com AUC= 0.56 e matriz de confusão na figura 8.7, com classificação dos pacientes vivos igual a 43, e 7 não classificados. Além de ter 2 classificações corretas e 2 não classificadas, para o estado do paciente morto.

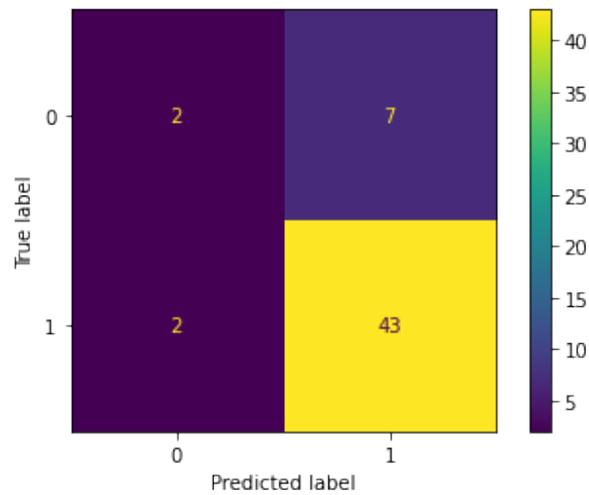
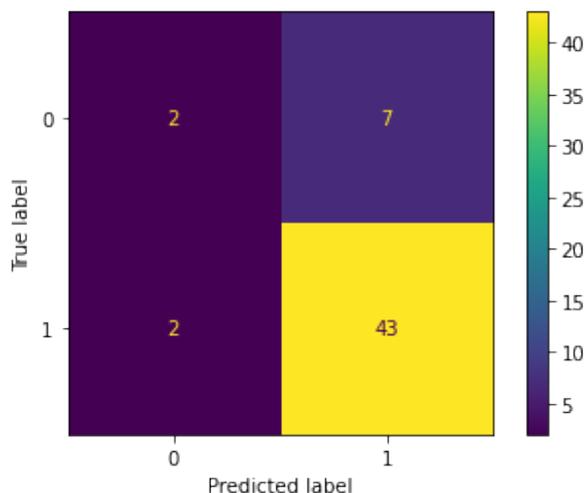


Figura 8.7: Matriz de confusão KNN - cenário 2

Figura 8.8: Matriz de Confusão SVM - cenário 2



8.2.3 SVM

Assim como os outros modelos, com o SVM não foi possível melhorar os resultados, pois obteve uma AUC = 0.64. Na matriz de confusão 8.8, foi possível obter 43 pacientes classificados, e 7 não classificados para o estado vivo, no entanto, 2 pacientes foram classificados e 2 não classificados para o estado de vida morto.

8.3 Cenário 3

Para este cenário, foi definido alguns critérios para ser considerado na elaboração dos modelos, pois consiste na predição e classificação dos pacientes conforme o tempo de internação. Foi efetuada uma conversão dos valores do tempo de internação, para uma forma binária, pela qual os pacientes que possuíam mais de 2880 minutos, equivalente a 48h, teriam valor 1, pacientes que tiveram tempo de internação maior a este período e os demais valores abaixo ou iguais a 48h, valor 0.

Além dessas conversões quanto ao tempo de internação obtidos, utilizou-se as variáveis probabilidade de morte e estado de vida do paciente para este cenário.

A separação binária dos pacientes a partir do tempo de internação, foi realizada a partir da hipótese de que as primeiras 24h não seriam consideradas, devido a necessidade de aplicação e avaliação da pontuação do SAPS, que de acordo com a literatura, obtém uma melhor eficácia nas primeiras 24h, mesmo que a base de dados possua pacientes com tempo de internação menor que 2880 minutos. A partir disso, foi considerado mais 24h, com total de 48h, para separar os pacientes da base de dados, e assim obter resultados que atendessem as hipóteses.

As variáveis utilizadas neste cenário, foram as mesmas do cenário 2, porém, a variável dependente é o tempo de internação do paciente na UTI, e as variáveis independentes são pro-

habilidade de morte e estado do paciente.

8.3.1 Regressão Logística

O algoritmo de regressão logística não obteve bons resultados, principalmente ao se tratar da acurácia com $AUC = 0.60$ - ver figura 8.12-, ao avaliar a sua curva ROC, não atingiu uma acurácia que tenha um nível de confiança equivalente aos outros modelos com AUC superior a 0.8.

Além disso, de acordo com a matriz de confusão do modelo presente na figura 8.9, foram classificados 45 pacientes e 9 não foram classificados com tempo de internação maior que 48h. E não obteve nenhuma classificação para pacientes com tempo de internação inferior a 48h.

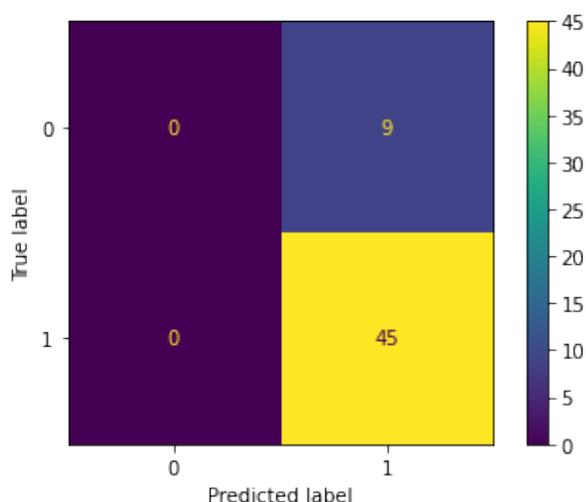


Figura 8.9: Matriz de confusão Regressão logística - cenário 3

8.3.2 KNN

Os resultados do algoritmo KNN foram mais satisfatórios, com $AUC = 0.67$, - ver figura 8.12-. Com isso, é possível obter um nível de confiança melhor quanto a sensibilidade e especificidade do algoritmo, isso, se comparado aos outros modelos deste cenário.

Ao avaliar a matriz de confusão presente na figura 8.10, nota-se que o algoritmo tem uma precisão maior quanto a pacientes que permanecerão na UTI por mais de 48h, com 43 classificados, e apenas 8 não. Isto posto, a quantidade de pacientes com valor 0 pelo modelo, estão muito abaixo se comparado a quantidade classificada com valor 1. Foram 2 classificados com tempo de internação inferior ou igual a 48h, 1 não classificado.

Figura 8.10: Matriz de Confusão KNN - cenário 3

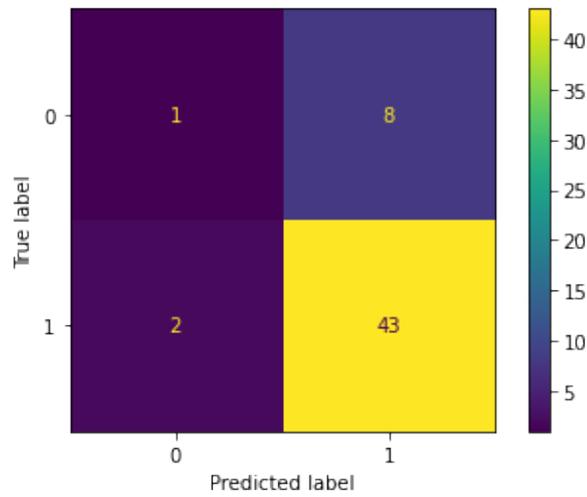


Figura 8.11: Matriz de Confusão SVM - cenário 3

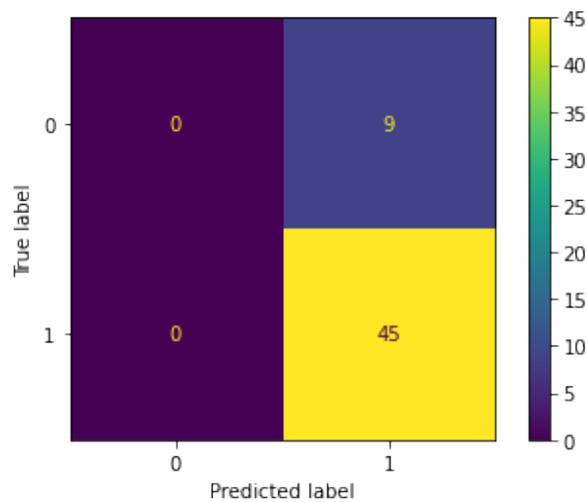
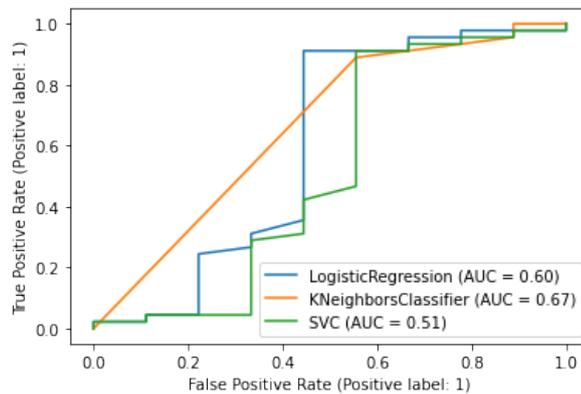


Figura 8.12: Curva ROC dos modelos implementados -cenário 3



8.3.3 SVM

Os resultados do algoritmo KNN foram mais satisfatórios, com $AUC = 0.67$, - ver figura 8.12-. Com isso, é possível obter um nível de confiança melhor quanto a sensibilidade e especificidade do algoritmo, isso, se comparado aos outros modelos deste cenário.

8.4 Cenário 4

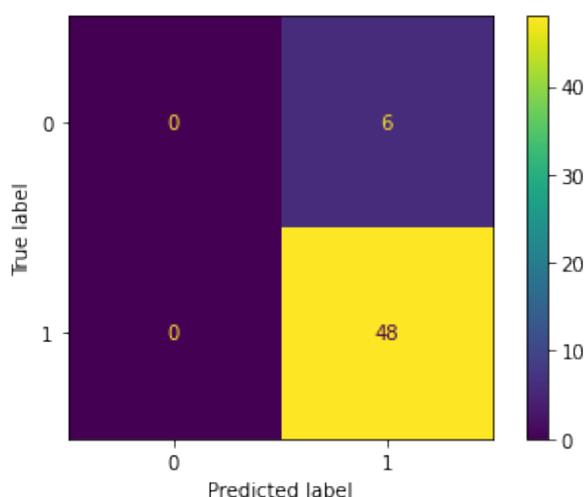
Na elaboração do cenário, foi considerado as mesmas variáveis utilizadas no cenário 3, com o objetivo de prever se o paciente irá permanecer internado por até 48h ou irá superar as 48h.

Ademais, neste cenário foi inferida a variável gênero, com o intuito de obter novos resultados e comparar os cenários, visto que com o cenário anterior, não foi possível superar a AUC dos modelos na literatura.

8.4.1 Regressão Logística

Com $AUC = 0.85$ de acordo com a curva ROC 8.16, obtendo o melhor resultado dentre os modelos propostos. Ainda assim, a matriz de confusão 8.13 não gerou uma predição para pacientes que permanecerão internados por até 48h, no entanto, foi obtida uma predição quanto aos pacientes que irão passar de 48h de internação, com 48 classificações e 6 não classificados.

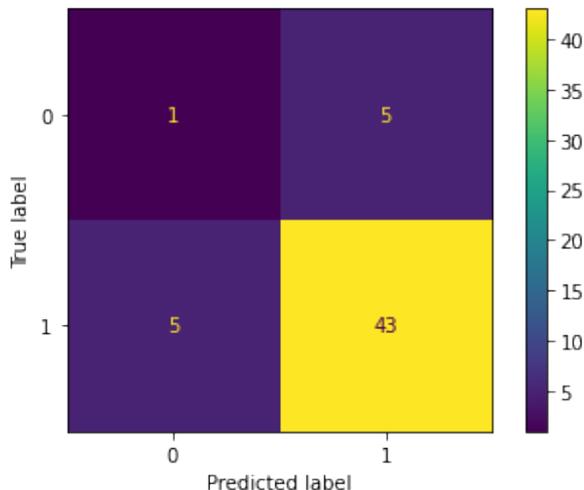
Figura 8.13: Matriz de confusão Regressão Logística -cenário 4



8.4.2 KNN

Já com o KNN, foi possível obter uma matriz de confusão mais acertiva quanto a predição, ao classificar 43 pacientes com probabilidade de tempo de internação maior que 48, e 5 pacientes com menos de 48h, por fim 6 não foram classificados, de acordo com a figura 8.14. No entanto, com $AUC = 0.64$, o modelo não superou a regressão logística neste cenário.

Figura 8.14: Matriz de confusão KNN -cenário 4



8.4.3 SVM

Por fim, o modelo que obteve os piores resultados nesse cenário, o SVM, possui uma matriz de confusão 8.15, que apesar de ter uma classificação parecida com a regressão logística, com predição de 48 pacientes com tempo de internação maior que 48h, e 6 não classificados, este modelo obteve AUC = 0.32.

Figura 8.15: Matriz de confusão SVM -cenário 4

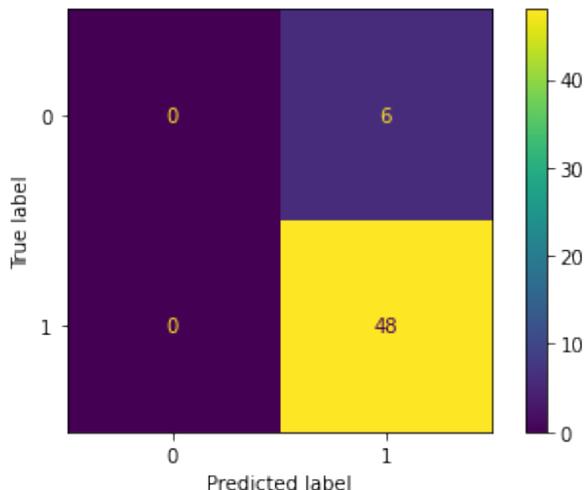
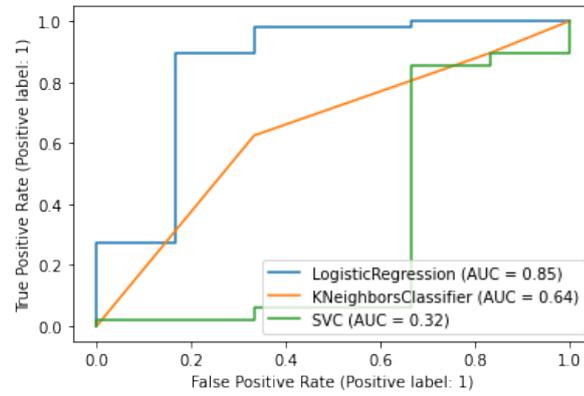


Figura 8.16: Curva ROC dos modelos implementados -cenário 4



Capítulo 9

Discussão

Esta seção descreve as técnicas que foram utilizadas para a obtenção dos resultados. Além disso, observa-se a comparação dos dados obtidos com relação aos modelos elaborados em cada cenário proposto anteriormente.

Sendo assim, a busca na literatura por dados relevantes, principalmente que atendessem às hipóteses criadas, foi dada através do uso de *strings* de busca cada vez mais específicas para outros tipos de doenças a fim de encontrar base de dados que publicam trabalhos científicos de boa relevância e disponibilizam os dados legendados, de modo que facilitem o uso das informações.

Portanto, a análise particular que fizemos foi de que o algoritmo de KNN obteve uma melhor validação, uma vez que por meio da variável foi possível obter maior acurácia no seu modelo com as variáveis de tempo de internação e a taxa de mortalidade calculada. Em contrapartida, o modelo de regressão logística teve uma acurácia menor, ainda que esse modelo tenha utilizado a variável gênero a mais no treinamento e teste.

Embora já citadas, as pontuações APACHE - IV, presentes na base de dados utilizada, não foram utilizadas. Com isso a metodologia APACHE, serviu apenas para comparar quantitativamente a AUC, o modelo com SVM, atingiu uma acurácia mais próxima do modelos APACHE, quanto ao prognóstico de mortalidade, com AUC= 0.85 no cenário 1, utilizando as variáveis gênero, probabilidade de morte calculada pelo SAPS-3 e tempo de internação do paciente.

Por conseguinte, foi obtido resultados quanto a predição do tempo de internação dos pacientes, ao considerar os intervalos, antes de 48h e depois de 48h. Com AUC = 0.85 t o modelo com Regressão Logística obteve o melhor resultado no cenário 4, ao utilizar as variáveis gênero, probabilidade de morte calculada pelo SAPS-3 e estado de vida do paciente.

- É possível obter prognóstico eficaz quanto a pacientes com septicemia.
- Aplicação de algoritmos de aprendizagem de máquina, obtém resultados iguais ou superiores de AUC, quanto aos métodos APACHE- IV e SAPS-3.

- Aplicação algoritmos de aprendizagem de máquina, predizem probabilidade de óbito de pacientes com septicemia.
- Aplicação algoritmos de aprendizagem de máquina, predizem probabilidade do tempo de intenação do paciente na UTI.

Idealmente a aplicação dos modelos devem ser validados posteriormente com médicos e especialistas no âmbito hospitalar, e assim obter uma resposta mais clara, quanto a eficácia dos modelos aqui criados, porém, não foi possível implementar um framework com interface mais acessível para os profissionais de saúde, e conseqüentemente estes modelos não foram disponibilizados para os profissionais ainda.

Capítulo 10

Conclusão

Houve uma dificuldade quanto a obtenção de dados reais e utilizáveis, assim, quando finalmente foi possível coletar os dados. Não foi possível perceber o tanto de dados duplicados que tinham na base de dados no primeiro momento, posteriormente após uma normalização, foi possível perceber pacientes com dados duplicados.

Quanto as hipóteses criadas no início da pesquisa, obteve respostas quanto a eficácia do prognóstico específico, com pacientes que foram acometidos com septicemia, além de ser possível aplicar modelos de aprendizagem de máquina que obtiveram AUC significativos, ao serem comparados com os modelos da literatura. Ademais, foi possível prever a probabilidade do estado de vida do paciente com sepse, ao utilizar o cálculo de probabilidade de morte do sistema SAPS, como parâmetro na implementação do modelo.

Quanto a última hipótese, foi possível prever se o tempo de internação do paciente na UTI, conforme os critérios adotados, para menor ou igual a 48h de tempo de internação, ou maior que 48h. Portanto, foi possível avaliar quantitativamente os modelos propostos por este trabalho. Porém, é perceptível que uma abordagem com maior número de pacientes, melhoraria consequentemente os resultados obtidos, ou pelo menos, traria mais conforto na análise dos modelos.

Para trabalhos futuros, esperamos poder contar com uma base de dados bem maior, para melhor treinar os algoritmos, assim poderemos chegar cada vez mais perto do modelo de predição ideal para prognóstico de pessoas com septicemia, e assim contribuir com a diminuição do tempo de internação dos pacientes, bem como ser um auxílio na tomada de decisão dos médicos e especialistas diminuindo as chances de morte das pessoas.

Referências

- Santos HG, nascimento CF, izbicki r, duarte YAO, chiavegatto filho ADP. machine learning para análises preditivas em saúde: exemplo de aplicação para predizer óbito em idosos de são paulo, brasil. *cad saúde pública* 2019 35(7):e00050818. *Cadernos de Saúde Pública*, 36(1), 2020. DOI 10.1590/0102-311xer050818. URL <https://doi.org/10.1590/0102-311xer050818>.
- Baran Balkan, Patrick Essay, and Vignesh Subbian. Evaluating icu clinical severity scoring systems and machine learning applications: Apache iv/iva case study. In *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 4073–4076, 2018. DOI 10.1109/EMBC.2018.8513324.
- Roger C Bone, Robert A Balk, Frank B Cerra, R Phillip Dellinger, Alan M Fein, William A Knaus, Roland M H Schein, and William J Sibbald. Definitions for sepsis and organ failure and guidelines for the use of innovative therapies in sepsis. *Chest*, 101(6):1644–1655, June 1992.
- Bernhard E. Boser, Isabelle M. Guyon, and Vladimir N. Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory, COLT '92*, page 144–152, New York, NY, USA, 1992. Association for Computing Machinery. ISBN 089791497X. DOI 10.1145/130385.130401. URL <https://doi.org/10.1145/130385.130401>.
- Andrew P. Bradley. The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30(7):1145–1159, 1997. ISSN 0031-3203. DOI [https://doi.org/10.1016/S0031-3203\(96\)00142-2](https://doi.org/10.1016/S0031-3203(96)00142-2). URL <https://www.sciencedirect.com/science/article/pii/S0031320396001422>.
- Steven C Brooks and Connie W Tsao. Circulation. <http://circ.ahajournals.org/content/101/23/e215.full>]; Accessed: 2022-7-16.
- Gyzelly Alves de Carvalho, Adriana Arruda Barbosa Rezende, Geovane Rossone Reis, and Giulliano Gardenghi. Uso do escore apache iv como preditor de mortalidade e tempo de permanência em uma unidade de terapia intensiva. *Revista Pesquisa em Fisioterapia*, 10(1):

- 9–15, jan. 2020. DOI [10.17267/2238-2704rpf.v10i1.2606](https://doi.org/10.17267/2238-2704rpf.v10i1.2606). URL <https://www5.bahiana.edu.br/index.php/fisioterapia/article/view/2606>.
- eICU Collaborative Research Database. eicu collaborative research database. Disponível em: <https://eicu-crd.mit.edu/eicutables/apacheapsvar/>. Acesso em: Fevereiro, 2022.
- F. Ellett, J. Jorgensen, Marand A. L., M. L. Yuki, M. M. Myrian, S. Vicky, Kathryn L. B., L. Jarone, and I. Daniel. Diagnosis of sepsis from a drop of blood by measurement of spontaneous neutrophil motility in a microfluidic assay. 2018. DOI <https://doi.org/10.1038/s41551-018-0208-z>.
- Framework para cálculo do APACHE IV. Framework para cálculo do apache iv. Disponível em: <https://intensivecarenetwork.com/Calculators/Files/Apache4.html>. Acesso em: Agosto, 2022.
- Mohammad Ghorbani, Haleh Ghaem, Abbas Rezaianzadeh, Zahra Shayan, Farid Zand, and Reza Nikandish. A study on the efficacy of APACHE-IV for predicting mortality and length of stay in an intensive care unit in iran. *F1000Res.*, 6:2032, November 2017.
- A L Goldberger, L A Amaral, L Glass, J M Hausdorff, P C Ivanov, R G Mark, J E Mietus, G B Moody, C K Peng, and H E Stanley. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. *Circulation*, 101(23):E215–20, June 2000.
- Chang Hu, Lu Li, Weipeng Huang, Tong Wu, Qiancheng Xu, Juan Liu, and Bo Hu. Interpretable machine learning for early prediction of prognosis in sepsis: A discovery and validation study. *Infectious Diseases and Therapy*, 11(3):1117–1132, April 2022. DOI [10.1007/s40121-022-00628-6](https://doi.org/10.1007/s40121-022-00628-6). URL <https://doi.org/10.1007/s40121-022-00628-6>.
- Knaus WA Draper EA Wagner DP Zimmerman JE. Apache ii: a severity of disease classification system. *Crit Care Med.*, 1985.
- Alistair Johnson, Tom Pollard, Omar Badawi, and Jesse Raffa. EICU collaborative research database demo, 2021.
- Alan E. Jones, Stephen Trzeciak, and Jeffrey A. Kline. The sequential organ failure assessment score for predicting outcome in patients with severe sepsis and evidence of hypoperfusion at the time of emergency department presentation*. *Critical Care Medicine*, 37(5) : 1649 – –1654, May 2009. DOI [10.1097/ccm.0b013e31819def97](https://doi.org/10.1097/ccm.0b013e31819def97). URL <https://doi.org/10.1097/ccm.0b013e31819def97>.
- V. N. Jordão, L. A. P. Nascimento, V. G. B. Lima, M. C. Farah, and H. P. Guimarães. Sepsis: a discussion on the changes of its diagnostic criteria. *Brazilian Journal of Health Review*, 2019.

- Simarjeet Kaur, Jimmy Singla, Lewis Nkenyereye, Sudan Jha, Deepak Prashar, Gyanendra Prasad Joshi, Shaker El-Sappagh, Md. Saiful Islam, and S. M. Riazul Islam. Medical diagnostic systems using artificial intelligence (ai) algorithms: Principles and perspectives. *IEEE Access*, 8:228049–228069, 2020. DOI [10.1109/ACCESS.2020.3042273](https://doi.org/10.1109/ACCESS.2020.3042273).
- HI Kim and S. Park. Sepsis: Early recognition and optimized treatment. *Epub*, 2019. DOI [10.4046/trd.2018.0041](https://doi.org/10.4046/trd.2018.0041).
- W A Knaus, J E Zimmerman, D P Wagner, E A Draper, and D E Lawrence. APACHE-acute physiology and chronic health evaluation: a physiologically based classification system. *Crit. Care Med.*, 9(8):591–597, August 1981.
- William A Knaus, Douglas P Wagner, Elizabeth A Draper, Jack E Zimmerman, Marilyn Bergner, Paulo G Bastos, Carl A Sirio, Donald J Murphy, Ted Lotring, Anne Damiano, and Frank E Harrell, Jr. The APACHE III prognostic system. *Chest*, 100(6):1619–1636, December 1991.
- Goh K.H. Wang L., Yeow A.Y.K., Hermione P., Ke L., Joannas J. L. Y., and Gamaliel Y. H. T. Artificial intelligence in sepsis early prediction and diagnosis using unstructured data in healthcare. 2021. DOI doi.org/10.1038/s41467-021-20910-4.
- M. M. Lavoisier, M. G. V. Ana, and U. A. Ulisses. Predictive model for sepsis in the intensive care unit. 2016. DOI [dx.doi.org/10.18569/tempus.v10i2.1832](https://doi.org/10.18569/tempus.v10i2.1832).
- Hui Li, Liang Liu, Dingyu Zhang, Jiuyang Xu, Huaping Dai, Nan Tang, Xiao Su, and Bin Cao. Sars-cov-2 and viral sepsis: observations and hypotheses. *The Lancet*, 395(10235):1517–1520, 2020. ISSN 0140-6736. DOI [https://doi.org/10.1016/S0140-6736\(20\)30920-X](https://doi.org/10.1016/S0140-6736(20)30920-X). URL <https://www.sciencedirect.com/science/article/pii/S014067362030920X>.
- S. M. Lobo, E. Rezende, C. L. Mendes, and M. C. Oliveira. Mortality due to sepsis in brazil in a real scenario: the brazilian icus project. mortalidade por sepse no brasil em um cenário real: projeto utis brasileiras. 2019. DOI doi.org/10.5935/0103-507X.20190008.
- Batta Mahesh. Machine learning algorithms-a review. *International Journal of Science and Research (IJSR)*. [Internet], 9:381–386, 2020.
- Qingqing Mao, Melissa Jay, Jana L Hoffman, Jacob Calvert, Christopher Barton, David Shimabukuro, Lisa Shieh, Uli Chettipally, Grant Fletcher, Yaniv Kerem, Yifan Zhou, and Ritanakar Das. Multicentre validation of a sepsis prediction algorithm using only vital sign data in the emergency department, general ward and icu. *BMJ Open*, 8(1), 2018. ISSN 2044-6055. DOI [10.1136/bmjopen-2017-017833](https://doi.org/10.1136/bmjopen-2017-017833). URL <https://bmjopen.bmj.com/content/8/1/e017833>.
- Marcos Vinícius. Modelos implementados no colab. Disponível em: <https://colab.research.google.com/drive/1qJYfMc740PquSTvqk3Yv-jATmcxZJcpE?usp=sharing>. Acesso em: Agosto, 2022a.

- Marcos Vinícius. Código github. Disponível em: <https://github.com/marcmec/PrognosticSepsis>. Acesso em: Agosto, 2022b.
- Marcos Vinícius. Código code ocean. Disponível em: <https://codeocean.com/capsule/1750428/tree>. Acesso em: Agosto, 2022c.
- Paul E. Marik and Abdalsamih M. Taeb. SIRS, qSOFA and new sepsis definition. *Journal of Thoracic Disease*, 9(4), 2017. ISSN 2077-6624. URL <https://jtd.amegroups.com/article/view/12738>.
- Ministério-da-Saúde. Conecte sus. Disponível em: <http://tabnet.datasus.gov.br/cgi/defctohtm.exe?sih/cnv/niuf.def>. Acesso em: Março, 2021.
- Senthilkumar Mohan, Chandrasegar Thirumalai, and Gautam Srivastava. Effective heart disease prediction using hybrid machine learning techniques. *IEEE Access*, 7:81542–81554, 2019. DOI [10.1109/ACCESS.2019.2923707](https://doi.org/10.1109/ACCESS.2019.2923707).
- Rui P Moreno, Philipp G H Metnitz, Eduardo Almeida, Barbara Jordan, Peter Bauer, Ricardo Abizanda Campos, Gaetano Iapichino, David Edbrooke, Maurizia Capuzzo, Jean-Roger Le Gall, and SAPS 3 Investigators. SAPS 3—from evaluation of the patient to evaluation of the intensive care unit. part 2: Development of a prognostic model for hospital mortality at ICU admission. *Intensive Care Med.*, 31(10):1345–1355, October 2005.
- F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.
- Tom J Pollard, Alistair E W Johnson, Jesse D Raffa, Leo A Celi, Roger G Mark, and Omar Badawi. The eICU collaborative research database, a freely available multi-center database for critical care research. *Scientific Data*, 5(1):180178, September 2018.
- D. Ren, C. Ren, Rq. Yao, Y. Feng, and Y. Yao. Clinical features and development of sepsis in patients infected with sars-cov-2: a retrospective analysis of 150 cases outside wuhan, china. *Intensive Care Med*, 2020. DOI <https://doi.org/10.1007/s00134-020-06084-5>.
- P. T. Santos. Estreptococcias - streptococcal infections. 1999. DOI [0021-7557/99/75-Supl.1/S103](https://doi.org/10.1007/978-95-001-1510-3_103).
- scikit-learn. Svm margins example. Disponível em: https://scikit-learn.org/stable/auto_examples/svm/. Acesso em: Agosto, 2022.
- Christopher W. Seymour, Foster Gesten, Hallie C. Prescott, Marcus E. Friedrich, Theodore J. Iwashyna, Gary S. Phillips, Stanley Lemeshow, Tiffany Osborn, Kathleen M. Terry, and Mitchell M. Levy. Time to treatment and mortality during mandated emergency care for sepsis. *New England Journal of Medicine*, 376(23):2235–2244, 2017. DOI [10.1056/NEJMoa1703058](https://doi.org/10.1056/NEJMoa1703058). URL <https://doi.org/10.1056/NEJMoa1703058>. PMID: 28528569.

João Manoel Silva Junior, Luiz M Sá Malbouisson, Hector L Nuevo, Luiz Gustavo T Barbosa, Lauro Yoiti Marubayashi, Isabel Cristina Teixeira, Antonio Paulo Nassar Junior, Maria Jose Carvalho Carmona, Israel Ferreira da Silva, José Otávio Costa Auler Júnior, and Ederlon Rezende. Applicability of the simplified acute physiology score (SAPS 3) in brazilian hospitals. *Rev. Bras. Anesthesiol.*, 60(1):20–31, January 2010.

Ishan Taneja, Gregory L. Damhorst, Carlos Lopez-Espina, Sihai Dave Zhao, Ruoqing Zhu, Shah Khan, Karen White, James Kumar, Andrew Vincent, Leon Yeh, Shirin Majdizadeh, William Weir, Scott Isbell, James Skinner, Manubolo Devanand, Syed Azharuddin, Rajamurugan Meenakshisundaram, Riddhi Upadhyay, Anwaruddin Syed, Thomas Bauman, Joseph Devito, Charles Heinzmann, Gregory Podolej, Lanxin Shen, Sanjay Sharma Timilsina, Lucas Quinlan, Setareh Manafirasi, Enrique Valera, Bobby Reddy, and Rashid Bashir. Diagnostic and prognostic capabilities of a biomarker and EMR-based machine learning algorithm for sepsis. *Clinical and Translational Science*, 14(4):1578–1589, May 2021. DOI [10.1111/cts.13030](https://doi.org/10.1111/cts.13030). URL <https://doi.org/10.1111/cts.13030>.

Kashvi Taunk, Sanjukta De, Srishti Verma, and Aleena Swetapadma. A brief review of nearest neighbor algorithm for learning and classification. In *2019 International Conference on Intelligent Computing and Control Systems (ICCS)*, pages 1255–1260, 2019. DOI [10.1109/ICCS45141.2019.9065747](https://doi.org/10.1109/ICCS45141.2019.9065747).

Franco van Wyk, Anahita Khojandi, Rishikesan Kamaleswaran, Oguz Akbilgic, Shamim Nemati, and Robert L. Davis. How much data should we collect? a case study in sepsis detection using deep learning. In *2017 IEEE Healthcare Innovations and Point of Care Technologies (HI-POCT)*, pages 109–112, 2017. DOI [10.1109/HIC.2017.8227596](https://doi.org/10.1109/HIC.2017.8227596).

Alfredo Vellido, Vicent Ribas, Carles Morales, Adolfo Ruiz Sanmartín, and Juan Carlos Ruiz Rodríguez. Machine learning in critical care: state-of-the-art and a sepsis case study. *BioMedical Engineering OnLine*, 17(S1), November 2018. DOI [10.1186/s12938-018-0569-2](https://doi.org/10.1186/s12938-018-0569-2). URL <https://doi.org/10.1186/s12938-018-0569-2>.

Achmad Widodo and Bo-Suk Yang. Support vector machine in machine condition monitoring and fault diagnosis. *Mechanical Systems and Signal Processing*, 21(6):2560–2574, 2007. ISSN 0888-3270. DOI <https://doi.org/10.1016/j.ymsp.2006.12.007>. URL <https://www.sciencedirect.com/science/article/pii/S0888327007000027>.

Qiangrong Zhai, Zi Lin, Hongxia Ge, Yang Liang, Nan Li, Qingbian Ma, and Chuyang Ye. Using machine learning tools to predict outcomes for emergency department intensive care unit patients. *Scientific Reports*, 10(1):20919, Dec 2020. ISSN 2045-2322. DOI [10.1038/s41598-020-77548-3](https://doi.org/10.1038/s41598-020-77548-3). URL <https://doi.org/10.1038/s41598-020-77548-3>.

Jack E Zimmerman, Andrew A Kramer, Douglas S McNair, and Fern M Malila. Acute physiology and chronic health evaluation (APACHE) IV: hospital mortality assessment for today's critically ill patients. *Crit. Care Med.*, 34(5):1297–1310, May 2006.

Franciele Nascimento Santos Zonta, Patricia Gurgel Amaral Velasquez, Leonardo Garcia Velasquez, Luciana Savi Demetrio, Danieli Miranda, and Mirian Bortolamedi Diamenes Silva. Epidemiological and clinical characteristics of sepsis in a public hospital of paran . *Revista de Epidemiologia e Controle de Infec o*, 8(3):224–231, 2018. ISSN 2238-3360. DOI 10.17058/reci.v8i3.11438. URL <https://online.unisc.br/seer/index.php/epidemiologia/article/view/11438>.

Apêndice A

Código para gerar os modelos de aprendizagem de máquina

```
1
2 # -*- coding: utf-8 -*-
3 """Visualizacao_de_dados
4
5 Automatically_generated_by_Colaboratory.
6
7 Original_file_is_located_at
8 https://colab.research.google.com/drive/1qJYfMc740PquSTvqk3Yv-
   jATmcxZJcpE
9 """
10
11 import pandas as pd
12 import io
13 from google.colab import files
14 import csv
15 import plotly.graph_objects as go
16 import numpy as np
17 from sklearn.datasets import make_moons
18 from sklearn.model_selection import train_test_split
19 from sklearn.neighbors import KNeighborsClassifier
20 import matplotlib.pyplot as plt
21 from math import e
22 from sklearn.linear_model import LogisticRegression
23 from sklearn import metrics
24 import seaborn as sn
25 import plotly.express as px
26 from sklearn.preprocessing import StandardScaler
```

```
27 from sklearn import svm
28 from sklearn.datasets import make_blobs, make_circles
29 from sklearn.metrics import classification_report, confusion_matrix,
    plot_confusion_matrix
30 from sklearn.model_selection import cross_val_score
31 from sklearn.model_selection import GridSearchCV
32
33 pd.set_option('display.max_rows', None)
34 #exibir todo o data set independente da quantidade de linhas
35
36 apacheAdmit = pd.read_csv('apacheMaisAdmitTime.csv')
37 #leitura do dataset normalizado com as variaveis que ser o
    utilizadas
38
39 # a seguir tratar mais uma vez os dados do dataframe lido.
40 #categorizar os pacientes ALIVE e EXPIRED para binario 1 ou 0 -
    como estado de vida do paciente VIVO OU MORTO
41 #Categorizar os pacientes de acordo com seu genero HOMEM E MULHER
    para binario 1 ou 0
42 #categorizar os pacientes que tem tempo de internacao maior que 48h
    (1 para maiores que 2880minutos e 0 para menores que 48h)
43
44 apacheAdmit = apacheAdmit.replace(to_replace = ['ALIVE','EXPIRED'],
    value = ['1','0'])
45
46 apacheAdmit = apacheAdmit.replace(to_replace = ['Male','Female'],
    value = ['1','0'])
47
48 #apacheAdmit = apacheAdmit.replace(to_replace = ['> 89'],value =
    ['90'])
49
50 apacheAdmit.loc[apacheAdmit['hospitaldischargeoffset'] > 2880, '
    hospitaldischargeoffset'] = '1'
51
52 apacheAdmit.loc[(apacheAdmit['hospitaldischargeoffset'] != '1'), '
    hospitaldischargeoffset'] = '0'
53
54 #calcular a probabilidade de morte de acordo com o saps 3 disponivel
    no dataframe
55
```

```
56 txLogit = -32.6659 + np.log(apacheAdmit['acutephysiologyscore']
    +20.5958)*7.3068
57
58 txMortality= e**txLogit/(1+e**txLogit)
59
60 #gerar a nova coluna do dataframe com o resultado da probabilidade de
    morte calculada
61
62 apacheAdmit.insert(7,'txMortality',txMortality)
63
64 #efetuar o treino de acordo com os cen rios
65
66 #cen rio para progn stico de mortalidade do paciente
67
68 # # X = apacheAdmit[["hospitaldischargeoffset", "gender", 'txMortality
    ']]
69 # # y = apacheAdmit['actualhospitalmortality']
70 # X = apacheAdmit[["hospitaldischargeoffset", 'txMortality']]
71 # y = apacheAdmit['actualhospitalmortality']
72
73 #cen rio para predicao e tempo de internacao do paciente
74 X = apacheAdmit[["actualhospitalmortality", "gender", 'txMortality']]
75 y = apacheAdmit['hospitaldischargeoffset']
76 # X = apacheAdmit[["actualhospitalmortality", 'txMortality']]
77 # y = apacheAdmit['hospitaldischargeoffset']
78
79
80 X_train, X_test, y_train, y_test = train_test_split(X, y, test_size
    =0.3)
81
82 #aplicar as bibliotecas com os algoritmos de aprendizagem de m quina
    REGRESSAO LOGISTICA, KNN E SVM
83
84 logistic_regression= LogisticRegression()
85 logistic_regression.fit(X_train,y_train)
86
87 logistic_regression= LogisticRegression()
88 logistic_regression.fit(X_train,y_train)
89 y_predrl=logistic_regression.predict(X_test)
90
```

```
91 # plotagem da matriz de confusao de acordo com a predicao do modelo
    de regressao logistica
92 metrics.ConfusionMatrixDisplay.from_predictions(y_test, y_predrl)
93 plt.show()
94
95 print('Accuracy:_', metrics.accuracy_score(y_test, y_predrl))
96 plt.show()
97
98 logistic_regression.score(X, y)
99
100 print(classification_report(y_test, y_predrl))
101
102 scaler = StandardScaler()
103 scaler.fit(X_train)
104
105 X_train = scaler.transform(X_train)
106 X_test = scaler.transform(X_test)
107
108 knn_class = KNeighborsClassifier(n_neighbors=3)
109 knn_class.fit(X_train, y_train)
110
111 y_pred_knn = knn_class.predict(X_test)
112
113 print(y_pred_knn)
114
115 knn_class.score(X, y)
116
117 # plotagem da matriz de confusao de acordo com a predicao do modelo
    do KNN
118 metrics.ConfusionMatrixDisplay.from_predictions(y_test, y_pred_knn)
119
120 plt.show()
121 print(classification_report(y_test, y_pred_knn))
122
123 accuracy = str(round(knn_class.score(X_test, y_test) * 100, 2))+"%"
124
125 print(accuracy)
126
127 clfSVM = svm.SVC(kernel='linear', random_state=None, C=1.0, cache_size
    =200, coef0=0.0)
128
```

```
129 clfSVM.fit(X_train, y_train)
130
131 y_predSVM = clfSVM.predict(X_test)
132
133 param_grid = {'C':[0.1,1,10,100,1000], 'gamma': [1,0.1,0.001,
            0.0001], 'kernel':['linear']}
134
135 grid = GridSearchCV(svm.SVC(),param_grid,refit=True,verbose=3)
136
137 grid.fit(X_train,y_train)
138
139 grid.best_params_
140
141 pred_post_grid = grid.predict(X_test)
142
143 print(classification_report(y_test,pred_post_grid))
144 print('/n')
145 print(confusion_matrix(y_test,pred_post_grid))
146
147 # plotagem da matriz de confusao de acordo com a predicao do modelo
    SVM
148 metrics.ConfusionMatrixDisplay.from_predictions(y_test, y_predSVM)
149
150 plt.show()
151 # print(classification_report(y_test,y_predSVM))
152
153 X_train[:,0]
154
155 #gerar Caracterstica de Opera o do Recepto ou CURVA ROC, dos
    modelos criados e comparar sua AUC para definir os melhores
    modelos
156
157 disp = metrics.plot_roc_curve(logistic_regression,X_test,y_test)
158 metrics.plot_roc_curve(knn_class,X_test,y_test,ax=disp.ax_)
159 metrics.plot_roc_curve(clfSVM,X_test,y_test, ax=disp.ax_)
```