Instituto Federal de Educação, Ciência e Tecnologia da Paraíba

Campus João Pessoa

Programa de Pós-Graduação em Tecnologia da Informação

Nível Mestrado Profissional

VICTOR MONTEIRO SILVA

PREVENDO O RISCO DE MORTALIDADE ENTRE IDOSOS INTERNADOS COM PNEUMONIA: UMA ABORDAGEM BASEADA EM APRENDIZADO DE MÁQUINA

DISSERTAÇÃO DE MESTRADO

JOÃO PESSOA

2022

VICTOR MONTEIRO SILVA

Prevendo o risco de mortalidade entre idosos internados com pneumonia: uma abordagem baseada em aprendizado de máquina

Dissertação apresentada como requisito parcial para obtenção do título de Mestre em Tecnologia da Informação, pelo Programa de Pós-Graduação em Tecnologia da Informação do Instituto Federal de Educação, Ciência e Tecnologia da Paraíba – IFPB.

Orientador: Profa. Dra. Damires Yluska de Souza Fernandes
Coorientador: Prof. Dr. Alex Sandro da Cunha Rêgo

João Pessoa
2022

Dados Internacionais de Catalogação na Publicação (CIP)
Biblioteca Nilo Peçanha do IFPB, *campus* João Pessoa

S586p Silva, Victor Monteiro.

Prevendo o risco de mortalidade entre idosos internados com pneumonia : uma abordagem baseada em aprendizado de máquina / Victor Monteiro Silva. – 2022.

58 f. : il.

Dissertação (Mestrado -Tecnologia da Informação) - Instituto Federal de Educação da Paraíba / Programa de Pós-Graduação em Tecnologia da Informação, 2022.

Orientação : Profª. D.ra Damires Yluska de S. Fernandes.

Coorientação : Profº D.r Alex Sandro da Cunha Rêgo.

1.Pneumonia adquirida na comunidade (PAC). 2. Risco de mortalidade. 3. Aprendizado de máquina. 4. Curva ROC. 5. Seleção de atributos. I. Título.

CDU 616.24-002(043)

Lucrecia Camilo de Lima
Bibliotecária – CRB 15/132

VICTOR MONTEIRO SILVA

Prevendo o risco de mortalidade entre idosos internados com pneumonia: uma abordagem baseada em aprendizado de máquina

Dissertação apresentada como requisito parcial para obtenção do título de Mestre em Tecnologia da Informação, pelo Programa de Pós-Graduação em Tecnologia da Informação do Instituto Federal de Educação, Ciência e Tecnologia da Paraíba – IFPB.

Aprovado em 18 de Março de 2022.

BANCA EXAMINADORA:

gov.br Documento assinado digitalmente
Thiago Jose Marques Moura
Data: 25/08/2022 11:53:02-0300
Verifique em https://verificador.iti.br

Prof. Dr. Thiago José Marques Moura – IFPB
Avaliador

gov.br Documento assinado digitalmente
CARLOS EDUARDO SANTOS PIRES
Data: 24/08/2022 10:31:44-0300
Verifique em https://verificador.iti.br

Prof. Dr. Carlos Eduardo Santos – UFCG
Avaliador Externo

gov.br Documento assinado digitalmente
DAMIRES YLUSKA DE SOUZA FERNANDES
Data: 11/11/2022 15:09:03-0300
Verifique em https://verificador.iti.br

Profa. Dra. Damires Yluska de Souza Fernandes (Orientador)
Prof. Dr. Alex Sandro da Cunha Rêgo (Coorientador)

Visto e permitida a impressão
João Pessoa

gov.br Documento assinado digitalmente
FRANCISCO PETRONIO ALENCAR DE MEDEI
Data: 11/11/2022 09:24:33-0300
Verifique em https://verificador.iti.br

Prof. Dr. Francisco Petrônio A. de Medeiros
Coordenador PPPGTI

Este trabalho é dedicado à minha amada esposa, que me apoia hoje e sempre.

AGRADECIMENTOS

Quero agradecer, em primeiro lugar a Deus que ilumina minha vida durante toda esta longa caminhada.
Agradeço a minha família e amigos, que, com muito carinho e apoio, não mediram esforços para que eu chegasse até esta etapa da minha vida.
Agradeço a minha esposa, cujo cuidado e dedicação foi que me deram a perseverança para seguir.
Agradeço por fim, especialmente, a meus orientadores, Professora Damires Yluska e Professor Alex Sandro, sem os quais, com seus ensinamentos e muita paciência, este trabalho não seria possível.

RESUMO

A Pneumonia Adquirida na Comunidade (PAC) é uma infecção respiratória grave que pode causar a perda de vida em pessoas de diferentes idades, especialmente em pacientes idosos hospitalizados. Em relação a essa faixa etária, as taxas de mortalidade por PAC podem chegar a 30% de todas as causas respiratórias de óbito. Essa dissertação de mestrado apresenta uma abordagem baseada em aprendizado de máquina para prever o risco de mortalidade de pacientes idosos internados com PAC. Esse documento tem como objetivo mostrar o processo metodológico e as contribuições obtidas. Algumas contribuições foram publicadas em artigos científicos que podem ser lidos por completo nos Apêndices A e B. A abordagem proposta é capaz não apenas de classificar pacientes em risco de mortalidade durante a internação, como também de estimar a probabilidade relativa à previsão. Também é apresentado neste documento uma análise acerca da importância de atributos para a classificação do modelo. Em termos de valor de Área sob Curva ROC (AUC), após o processo de redução de dimensionalidade, o modelo Regressão Logística conseguiu um resultado de 0.86. O modelo treinado também consegue estimar probabilidades de classificação positiva que variam de 50 a 99%, ou seja, pacientes que podem vir a óbito. Acreditamos que a abordagem criada pode auxiliar a equipe médica a identificar com antecedência pacientes em iminência de estado crítico, de modo a melhorar a qualidade do tratamento e aumentar suas chances de recuperação.

Palavras-chaves: PAC. Risco de Mortalidade. Aprendizado de Máquina. AUC. Curva ROC. Seleção de atributos.

ABSTRACT

Community-Acquired Pneumonia (CAP) is a serious respiratory infection that can cause death in people of different ages, especially in hospitalized elderly patients. In this age group, mortality rates from CAP can reach 30% of all respiratory causes of death. This master's thesis presents a machine learning-based approach to predict the mortality risk of elderly patients hospitalized with CAP. This document includes the methodological process accomplished to develop the work along with the contributions obtained. Some contributions have been published by means of papers which have been added as appendices in this document. The proposed approach is able not only to classify patients at risk of mortality during hospitalization, but also to estimate the probability relative to prediction. An analysis regarding the importance of features for model classification is also presented in this document. In terms of Area Under ROC Curve (AUC) value, after applying feature selection techniques, the Logistic Regression model achieved a result of 0.86. The trained model is also able to estimate probabilities of positive classification ranging from 50 to 99%, i.e., patients who may die. We believe that the presented approach can help medical staff to identify patients in imminent critical condition in advance. Thus it can improve patients quality of treatment and increase their chances of recovery.

Key-words: CAP. Mortality Risk. Machine Learning. AUC. ROC Curve. Feature Selection.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ANOVA	Analysis Of Variance
AM	Aprendizado de Máquina
AUC	Area Under the Curve
CRISP-DM	Cross Industry Standard Process for Data Mining
CURB-65	Confusion, Uremia, Respiratory rate, Blood Pressure, Age > 65 years
MD	Mineração de Dados
LR	Logistic Regression
MLP	Multi-Layer Perceptron
PAC	Pneumonia Adquirida na Comunidade
PPGTI	Programa de Pós-Graduação em Tecnologia da Informação
PSI	Pneumonia Severity Index
RF	Random Forest
RME	Registro Médico Eletrônico
RNA	Rede Neural Artificial
ROC	Receiver Operating Characteristic
SVM	Support Vector Machine

SUMÁRIO

1 INTRODUÇÃO

A Pneumonia Adquirida na Comunidade (PAC) é uma infecção respiratória grave que pode causar risco de vida em pessoas de diferentes idades (World Health Organization, 2015). Sendo uma das infecções mais comuns que resultam na necessidade de hospitalização, a PAC pode inflamar os alvéolos pulmonares de um ou ambos os pulmões, afetar outros órgãos vitais e causar dificuldade para respirar. Casos de PAC são considerados difíceis de lidar, pois existem chances consideráveis que existam complicações se o paciente for um idoso, uma criança muito jovem, se tiver um sistema imunológico debilitado ou um problema médico sério como diabetes ou cirrose (WU et al., 2019).

Com o progresso da ciência médica, o acesso aos cuidados de saúde tem sido cada vez maior. Existem unidades especializadas com sofisticados sistemas de suporte à vida. Apesar disso, as taxas de mortalidade por PAC ainda podem chegar a 30% de todas as causas respiratórias de morte, principalmente no que diz respeito a pacientes idosos internados (HESPANHOL; BáRBARA, 2020). De fato, a PAC pode ser particularmente grave em pessoas com 65 anos ou mais, implicando em maior risco de mortalidade quando comparado a outras faixas etárias.

Para auxiliar na tomada de decisão no que concerne ao tratamento de pneumonia, dois escores médicos são comumente utilizados por profissionais de saúde, a saber (LONG; LONG; KOYFMAN, 2017): *Pneumonia Severity Index* (PSI) e *Confusion, Uremia, Respiratory rate, Blood Pressure* (*CURB-65*) e idade de 65 anos ou mais). Ambos os escores atuam como um método preliminar para prognósticos de mortalidade em pacientes internados com PAC, fornecendo à equipe médica um alerta com base nos dados do Registro Médico Eletrônico (RME) do paciente (RYAN et al., 2020). No entanto, esses escores médicos carecem de eficiência para respaldar essa predição, fornecendo apenas uma estimativa de no máximo 27% de confiança no indicativo de possibilidade do paciente vir a óbito, no caso do CURB-65, e até 29% no caso do PSI. Isso pode ser justificado devido ao fato de que os resultados do escore consideram apenas o estado atual de alguns dados do RME de um determinado paciente, excluindo a própria evolução do tratamento e outras medições clínicas relacionadas. Também não leva em consideração outros exemplos de pacientes com condições similares em termos de sintomas gerais, sinais, prognósticos e progressões (WIEMKEN; KELLEY; RAMIREZ, 2013).

Com base nesse contexto, esse trabalho de mestrado propõe uma abordagem computacional que objetiva mitigar os seguintes problemas:

(i) Como identificar pacientes idosos internados e diagnosticados com PAC em risco de falecerem durante o tratamento?

(ii) Como fornecer a probabilidade de que tal previsão possa de fato ocorrer?

(iii) Quais os atributos mais importantes para se obter uma predição de risco de falecimento que seja mais assertiva durante o tratamento?

Para responder às questões definidas e propor uma abordagem computacional, as seguintes etapas foram percorridas neste trabalho: (a) realização de um levantamento do estado da arte inicialmente feito por meio de uma revisão sistemática da literatura (SILVA et al., 2020); (b) definição e desenvolvimento da abordagem para identificação de pacientes com a probabilidade de risco de óbito e sua avaliação experimental (SILVA; SOUZA; RÊGO, 2022) e, por fim, (c) um estudo sobre os atributos mais relevantes para o problema de classificação em questão.

A realização das etapas (a) e (b) derivaram em artigos publicados em conferências com qualis em Computação. A revisão sistemática da literatura foi publicada nos anais do Symposium on Knowledge Discovery, Mining and Learning (KDMiLe) 2020 e a abordagem, contribuição principal desta pesquisa, foi publicada nos anais do International Conference on Enterprise Information Systems (ICEIS) 2022.

Além das publicações, o trabalho desta pesquisa derivou em um registro de software com o título: Hospital Modelo - Painel de Classificação de Risco de Mortalidade para Pacientes com Pneumonia, sob código de registro BR512022000996-0.

Este documento apresenta a dissertação resultante deste trabalho de mestrado, entretanto, sua organização difere dos moldes tradicionais. A razão para isso é que o Regulamento do Programa de Pós-Graduação em Tecnologia da Informação (PPGTI) inclui a possibilidade de defesa de mestrado por meio da aprovação de artigo com Qualis restrito e pedido de registro de software, ambos associados às contribuições do trabalho do mestrado. Tendo em vista que os pré-requisitos necessários para este formato de trabalho de conclusão de curso de mestrado foram alcançados no decorrer desta pesquisa, este documento foi estruturado de maneira diferenciada e está organizado conforme os seguintes capítulos, a saber:

(A) Introdução, com a motivação, problema de pesquisa e solução proposta pelo trabalho;

(B) Visão geral da revisão sistemática da literatura que baseou o trabalho de pesquisa;

(C) Visão geral da abordagem proposta para classificação de pacientes diagnosticados com PAC em risco de morte;

(D) Estudo para identificação dos atributos mais relevantes para o problema de classificação;

(E) Considerações e trabalhos futuros, com as principais contribuições e limitações do trabalho;

(F) Apêndice A com o artigo publicado com a revisão sistemática de literatura (SILVA et al., 2020);

(G) Apêndice B com o artigo publicado com a abordagem proposta para classificação de pacientes diagnosticados com PAC em risco de morte (SILVA; SOUZA; RÊGO, 2022).

2 REVISÃO SISTEMÁTICA DE LITERATURA

A Revisão Sistemática da Literatura (RSL), retratada no Apêndice A, apresenta um panorama acerca do estado da arte da aplicação de mineração de dados e aprendizado de máquina com vistas à assistência à tomada de decisões em tratamentos de pneumonia. O artigo resultante deste trabalho foi apresentado e publicado no *Symposium on Knowledge Discovery, Mining and Learning* (KDMiLe) no ano de 2020. Buscou-se, por meio deste estudo, mapear as evidências metodológicas que concernem a convergência das áreas de aprendizado de máquina e suporte à decisão clínica para tratamento de pneumonia, destacando lacunas e oportunidades de aprimoramentos nas pesquisas. Objetivou-se utilizar os resultados obtidos com o levantamento para basear a construção da abordagem de classificação, pauta deste trabalho de mestrado.

A principal questão de pesquisa tratada na RSL foi: Como a Mineração de Dados (MD) e o Aprendizado de Máquina (AM) estão sendo aplicados no apoio à tomada de decisão clínica em quadros de pneumonia? Essa questão foi decomposta em questões específicas de pesquisa, conforme descrição seguinte:

- QP1: Quais categorias de assistência à tomada de decisão em quadros de pneumonia são alvo das pesquisas?
- QP2: Quais tarefas de MD foram identificadas para as categorias de assistência?
- QP3: Quais modelos de AM encontrados para cada tarefa?
- QP4: Quais métricas utilizadas para avaliação dos modelos?

A seleção de estudos foi realizada por meio da seguinte string geral de busca: ("*machine learning*"OR "*data mining*"OR "*deep learning*") AND ("pneumonia")). Na primeira fase da RSL, foram obtidos 563 resultados. Em seguida foi realizada a leitura do título e do resumo dos artigos e, posteriormente, foi realizada a leitura das seções de introdução e conclusão dos artigos.

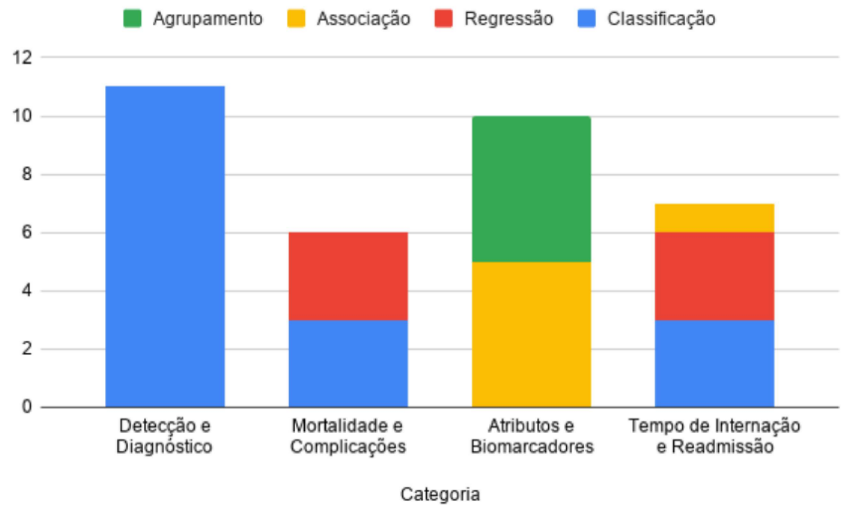
Os critérios de inclusão utilizados como filtros em cada uma dessas etapas foram: (i) estudos que apresentem a aplicação de técnicas de MD em tratamento ou prevenção de pneumonia; (ii) estudos que respondem a pelo menos uma das questões de pesquisa específicas.

Também foram verificados critérios de exclusão, a saber: (i) estudos que não proveem relevância científica; (ii) artigos publicados antes de 2010; (iii) estudos secundários ou terciários; (iv) estudos sem experimentação ou avaliação; (v) estudos que não consideram dados estruturados ou semi-estruturados.

Após a análise realizada em cada uma das etapas, levando em consideração os critérios de inclusão e exclusão, resultou-se em um total de 34 estudos que foram lidos por completo.

A partir dos artigos selecionados, foi realizada uma categorização desses trabalhos, destacando diferentes aplicabilidades de pesquisas no tratamento de pneumonia a saber: (i) Trabalhos com foco em detecção e diagnóstico de pneumonia; (ii) Trabalhos que objetivaram sinalizar risco de complicações em pacientes; (iii) Trabalhos que exploraram a análise de atributos e biomarcadores e (iv) Trabalhos com foco em previsão de tempo de internação e readmissão de pacientes. O gráfico ilustrado na Figura 1 sumariza a parcela de ocorrência das tarefas clássicas de MD por categoria de trabalhos analisados. Observa-se que a tarefa de classificação está presente na maioria das espécies de trabalhos, com predominância em estudos relacionados à detecção e diagnóstico de pneumonia.

Figura 1 – Tarefas de Mineração de Dados por Categoria.



Fonte: Próprio Autor

Posteriormente foram destacados os algoritmos de aprendizado mais utilizados em cada uma das tarefas de AM. Como pode ser visto na Tabela 1, o algoritmo *Support Vector Machine* (SVM) foi identificado como o mais utilizado na maioria dos trabalhos associados à predição de quadros de pneumonia. Os algoritmos *Logistic Regression* (LR) e *Random Forest* (RF) também ficaram entre os três modelos mais utilizados.

Tabela 1 – Quantitativos dos modelos de mineração

Modelos	**Quantitativo**
Support Vector Machine (SVM)	13
Rede Neural Artifical(RNA)	6
Random Forest (RF)	9
Logistic Regression (LR)	9
Árvore de Decisão (AD)	6
Naive Bayes (NB)	3
Total	**46**

Por fim, foi realizada uma sondagem quanto às medidas e/ou métricas utilizadas para avaliar o desempenho dos modelos de AM, cujo resultado está evidenciado na Tabela 2. Na maioria dos resultados, a plotagem gráfica da curva ROC foi a escolha mais empregada para analisar visualmente o desempenho dos modelos de classificação binário e, consequentemente, a Área sob a Curva (AUC) como medida que numericamente sumariza o desempenho do classificador a partir de uma curva ROC.

Tabela 2 – Principais métricas de avaliação por modelo

Modelos	**AUC**	**Acurácia**	**Sensitividade**	**Especificidade**	**Score kappa**
SVM	10	8	9	8	2
RNA	6	4	4	3	1
RF	9	5	6	6	2
LR	9	3	4	4	0
AD	5	4	3	3	3
NB	3	2	2	2	2
Total	**42**	**26**	**28**	**26**	**10**

Os resultados obtidos na revisão subsidiaram a escolha dos modelos de classificação e os métodos e métricas de avaliação da abordagem proposta, como descrita no Capítulo 3.

3 ABORDAGEM PARA CLASSIFICAÇÃO DE PACIENTES EM RISCO DE ÓBITO TRATAMENTO

Tendo em consideração os resultados obtidos na RSL e uma nova revisão ad-hoc com respeito a outros trabalhos relacionados, foi especificada a abordagem de classificação para identificar pacientes em risco de óbito por PAC. Este capítulo provê uma visão geral da abordagem.

3.1 Visão geral da abordagem

Dado o contexto do problema, duas questões de pesquisa foram definidas para guiar o estudo acerca da construção da abordagem: (i) Como identificar os pacientes idosos internados, diagnosticados com PAC, em risco de morte? E (ii) como fornecer a probabilidade de que tal previsão possa realmente ocorrer?

Com base nessas questões de pesquisa, a abordagem baseada em aprendizado de máquina foi definida e utilizou como base metodológica o modelo de processo *Cross Industry Standard Process for Data Mining*(CRISP-DM) (SCHRÖER; KRUSE; GÓMEZ, 2021). A abordagem provê um modelo para analisar e prever o risco de morte de pacientes idosos internados com PAC.

O problema de predição de risco de mortalidade pode ser compreendido como um problema de classificação binária. A classe positiva representa o risco de um paciente idoso internado vir a óbito durante a hospitalização, e a classe negativa indica a ausência deste risco. O classificador também faz a estimativa da probabilidade de confiança da predição para a classe positiva.

Painéis de acompanhamento de pacientes, comuns na maioria dos hospitais, sinalizam por meio de escores médicos o risco de mortalidade de pacientes. Seguindo essa logística, o CURB-65 foi utilizado duplamente neste trabalho para: (i) Definir quais atributos seriam incluídos no conjunto de dados de entrada e (ii) Servir de *baseline* conforme somatória dos critérios posisitovs estabelecida na literatura para classificação de risco (REF).

3.2 Conjunto de Dados

Para o problema de predição em questão, com base nos atributos necessários ao cálculo do CURB-65, 29 atributos foram selecionados como relevantes. Atributos numéricos incluem: idade, tempo de internação (medido em horas), pulso, frequência respiratória, pressão arterial sistólica, pressão arterial diastólica, temperatura, ureia, sódio, glicose e hematócrito. Considerando que nem toda aferição incluía todos os valores numéricos no conjunto de dados, foram

identificados valores nulos, os quais foram preenchidos utilizando o valor de mediana da aferição correspondente.

Os atributos categóricos representam a presença ou ausência de uma determinada condição no paciente. São eles: Morador de asilo, histórico de tabagismo, estado mental alterado, ventilação mecânica, doença neoplásica, insuficiência cardíaca congestiva, doença cerebrovascular, doença renal, doença hepática, doença pulmonar cronica, doença cardiovascular, doença psiquiátrica e doença neurológica.

Da mesma forma, um atributo de histórico de saúde familiar traz casos de doenças que também podem ser relevantes para compreender o diagnóstico e a evolução de um paciente (por exemplo, uma doença neurológica). O sexo do paciente é um atributo categórico, mas a sua classificação é 0 para os pacientes do sexo masculino e 1 para os do sexo feminino, como meio de padronização para ser incluso como parâmetro nos modelos que não aceitam valores não-numéricos.

Atributos referentes a comorbidades ou histórico familiar dos pacientes foram extraídos das anotações médicas e de enfermagem dos pacientes. Por exemplo, se uma anotação de um paciente contém o termo "Doença Pulmonar Crônica", o atributo de mesmo nome é definido como 1, ou 0 caso contrário. Uma visão dos atributos por categoria é apresentada no Apêndice B, Tabela 1.

Os rótulos das classes foram derivados de acordo com a seguinte lógica: exemplos de pacientes que têm informações sobre o tempo e a causa da sua morte relacionada com a PAC foram rotulados com 1 (falecido, 43% dos registros). Os pacientes que permaneceram vivos depois da hospitalização foram rotulados como 0 (sobrevivente, 57% dos registros).

3.3 Experimentos

Os experimentos foram planejados com o objetivo de avaliar a abordagem na tarefa de classificar pacientes em risco de falecerem. Os modelos foram avaliados utilizando a técnica de *10-fold grouped cross-validation* (HASTIE; TIBSHIRANI; FRIEDMAN, 2017). Os algoritmos de classificação utilizados foram o Random Forest (RF), Support Vector Machine (SVM), Multi-Layer Perceptron (MLP) e Logistic Regression (LR), por serem os mais utilizados para predições na área de pneumonia, com base na RSL.

O *baseline* para comparação com este trabalho foi desenvolvido como se segue: para um determinado vetor de atributos do paciente foi calculada a pontuação do CURB-65, atribuindo uma classificação positiva se o cálculo fosse maior que 3. Este valor representa um risco de mortalidade grave pela escala do CURB-65. A saída fornecida pelo CURB-65 também estima a probabilidade de mortalidade risco (até 27,8%).

A avaliação experimental incluiu a realização das seguintes atividades: (i) extração de

dados e de atributos específicos de anotações médicas e de enfermagem; (ii) definição de *baseline* desenvolvida de acordo com um escore real utilizado em hospitais; (iii) avaliação do desempenho do modelo preditor em comparação com o *baseline*, sob a perspectiva da métrica AUC e análise de curva ROC, e (iv) um teste de significância estatística para avaliar se o desempenho do classificador é superior ao apresentado pelo *baseline*.

3.4 Resultados

A Tabela 3 apresenta o resultado da AUC e a probabilidade de confiança da predição obtidos pelos classificadores examinados e o *baseline*. O classificador LR apresentou a maior AUC (0.81) para predição do risco de mortalidade, proporcionando uma média de probabilidade de classificação da classe positiva de 78%. O *baseline* alcançou uma AUC de 0.61, com uma média de probabilidade de 20%.

Tabela 3 – Resultados da experimentação considerando AUC e probabilidade

Modelo	**AUC**	**Probabilidade(%)**		
		MIN	**MAX**	**MEDIA**
Baseline	0.61	14	27	20
RF	0.78	50	92	68
SVM	0.71	50	86	75
MLP	0.75	50	98	78
LR	0.81	50	99	78

O teste de significância realizado confirmou que o desempenho da abordagem é estatisticamente superior ao baseline, com um nível de confiança de 95%.

Também foi avaliada uma análise da cronologia de cada aferição de determinados pacientes do conjunto de testes (Apêndice B, Tabela 3). Nos primeiros registros de aferição é possível verificar uma probabilidade baixa de risco de morte. Após 30 dias de internação, o classificador passa a indicar uma classificação positiva na maioria dos casos. A variação de atributos numéricos das medições fora dos seus valores de normalidade também influenciaram uma classificação positiva. Como exemplo, o atributo pulso, que normalmente varia de 57 a 100 em idosos, ao atingir um valor 140 influenciou o classificador a determinar que o paciente estava em risco de morte. A utilização de ventilação mecânica também indica uma probabilidade de morte cada vez maior. Esse impacto dos atributos que influenciaram na classificação dos modelos introduziu a necessidade de uma análise mais aprofundada, o qual é descrita no próximo capítulo.

O Apêndice B apresenta a abordagem de predição de risco de mortes de pacientes idosos por PAC em detalhes por meio do artigo publicado na conferência ICEIS 2022.

4 SELEÇÃO DE ATRIBUTOS PARA OTIMIZAÇÃO DO DESEMPENHO DO CLASSIFICADOR

Neste capítulo, é abordado o resultado do estudo acerca da seleção dos atributos mais relevantes para a melhoria do desempenho do classificador apresentado no capítulo anterior. Para isso, são introduzidos os conceitos sobre as técnicas de seleção de atributos exploradas. Também são discutidos alguns trabalhos relacionados que utilizaram métodos de seleção de atributos no contexto da área de saúde. Por fim, são descritos os cenários de experimentação para este estudo e os resultados obtidos.

4.1 Fundamentação teórica e trabalhos relacionados

Esta seção introduz conceitos e trabalhos relacionados a seleção de atributos associados a problemas de predição de risco de complicações em pacientes com pneumonia e outras infecções respiratórias.

4.1.1 Introdução à seleção de atributos

Em processos de análise preditiva, é natural que certos conjuntos de dados possam apresentar muitos atributos. Os atributos influenciam diretamente nos modelos preditivos e nos resultados que se pode alcançar. A quantidade de atributos pode levar a uma alta dimensão que, em algumas situações, pode implicar na prevalência de dados ruidosos, irrelevantes e/ou redundantes. Na prática, alguns atributos adicionam pouca informação ao objetivo da classificação e impacta no custo computacional para a construção do modelo preditor de acordo com o algoritmo de AM utilizado. Além disso, há a possibilidade de até mesmo interferir na criação de um bom modelo preditor.

Para tratar aspectos associados a esse contexto, técnicas como seleção e extração de atributos são comumente utilizadas como abordagens de redução de dimensionalidade de dados (VENKATESH; ANURADHA, 2019).

Métodos de seleção de atributos são utilizados para identificar os atributos mais relevantes para o modelo preditor com base no problema que se deseja resolver , enquanto métodos de extração de atributos, por outro lado, focam em transformar dados brutos em combinações de atributos que podem ser processados, preservando as informações no conjunto de dados original (VENKATESH; ANURADHA, 2019).

Um subconjunto dos atributos mais relevantes para uso na construção de um modelo de aprendizado de máquina melhora a precisão preditiva, aumenta a compreensibilidade dos resultados e reduz o tamanho do conjunto de dados (BELLMAN, 1966)(VENKATESH; ANURADHA,

2019).

O processo de seleção de atributos tem sido aplicado em uma ampla gama de problemas que incluem processamento de dados biológicos, finanças e sistemas de detecção de intrusão. Em particular, a seleção de atributos tem sido usada com sucesso em aplicações médicas, onde pode não apenas reduzir a dimensionalidade, mas também auxiliar a entender o comportamento e relação de certos atributos com a classe do problema de uma doença (REMESEIRO; BOLON-CANEDO, 2019).

4.1.2 Métodos para seleção de atributos

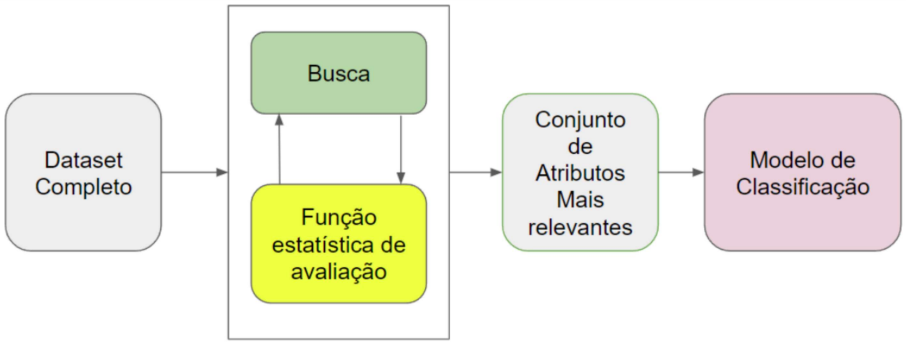
Existe uma variedade de métodos de seleção de atributos disponíveis no estado da arte. Para identificar atributos relevantes e melhorar o desempenho de classificadores, este trabalho considerou os três métodos comumente mais aplicados, a saber (DASH; LIU, 1997)(AGGARWAL; BALI; MITTAL, 2019): *Filter*; *Wrapper* e *Embedded*.

O método *Filter* analisa os atributos do conjunto de dados independente do algoritmo de aprendizado de máquina a ser treinado. Nesses métodos, os atributos são selecionados com base na aplicação de medidas estatísticas tais como ganho de informação, teste qui-quadrado(*chi-square*), pontuação de Fisher, coeficiente de correlação de pearson e análise de variância(ANOVA) (DASH;LIU, 1997). O método *filter*, na prática, requer menos tempo computacional pois não depende do processamento do modelo preditivo haja vista que se baseia em critérios discriminativos para determinação da classe do problema, tornando-o independente de qualquer algoritmo em particular.

A Figura 2 mostra uma visão geral do funcionamento de um método *filter*. Para um dataset de entrada, seus atributos são examinados conforme a aplicação de uma função estatística específica, de maneira que sejam identificados os considerados "mais relevantes" por meio de um ranqueamento obtido conforme função estatística empregada. Estes atributos são usados, então, para treinar o modelo de classificação. No método estatístico ANOVA, por exemplo, a função *F-Value* determina a importância de um atributo, a função é calculada com base na relação entre a variância do atributo entre grupos, dividido pela variância dentro dos grupos.

O método *wrapper* possui um princípio metodológico diferente em relação ao *filter*. Em vez de aplicar uma função estatística para determinar a importância do atributo, o algoritmo de aprendizado é treinado a partir de combinações distintas de atributos, com o intuito de identificar o subconjunto que maximiza o desempenho do classificador (AGGARWAL; BALI; MITTAL, 2019). Logo, o melhor subconjunto de atributos é dependente do classificador utilizado. O método *wrapper* é computacionalmente mais custoso do que o método *filter*, devido às repetidas etapas de aprendizado, utilizando técnicas como validação cruzada. No entanto, geralmente a utilização do método *wrapper* para seleção de atributos consegue alcançar melhor resultado na avaliação (AGGARWAL; BALI; MITTAL, 2019).

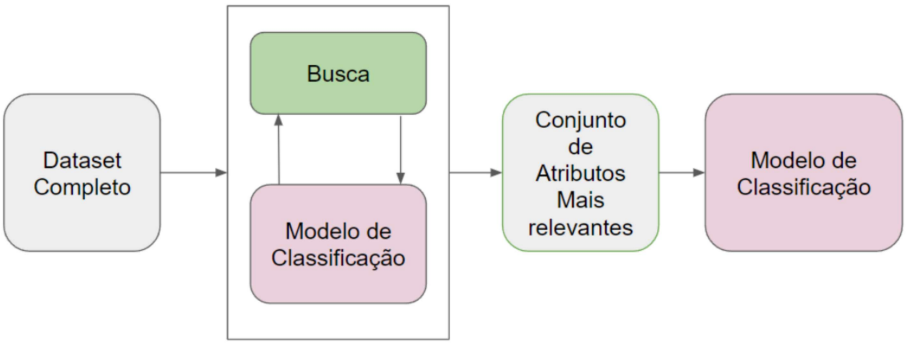
Figura 2 – Visão geral do funcionamento do método Filter



Fonte: Adaptado de (DASH; LIU, 1997)

A Figura 3 mostra uma visão geral do funcionamento do método *wrapper*. Alguns dos algoritmos empregados no método *Wrapper* são (AGGARWAL; BALI; MITTAL, 2019): *Recursive Feature Elimination* (RFE); Algoritmos de seleção de atributos sequenciais e Algoritmos genéticos. Em particular, após o primeiro treino, o algoritmo RFE irá verificar a importância dos atributos métodos estatísticos, então, recursivamente, irá remover os atributos menos importantes do *dataset* e treinar o modelo novamente. Por fim, será devolvido o conjunto de atributos mais relevantes para o modelo de classificação, após as iterações ocorridas de aprendizado e avaliação de atributos (PROVOST, 2000).

Figura 3 – Visão geral do funcionamento do método *Wrapper*



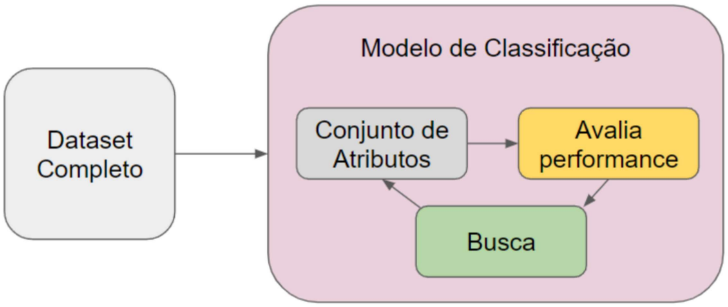
Fonte: Adaptado de (DASH; LIU, 1997)

As estratégias do tipo *embedded* são diretamente incorporadas ao algoritmo responsável pela indução do modelo preditivo (DASH; LIU, 1997). Os métodos *embedded* selecionam o subconjunto de atributos no próprio processo de construção do modelo, durante a fase de treinamento, e são geralmente específicos para um dado algoritmo de classificação (AGGARWAL; BALI; MITTAL, 2019).

A Figura 4 mostra o funcionamento de um método *embedded* onde, durante o próprio treinamento do modelo de classificação, são criados e avaliados conjuntos de atributos. Nessas etapas, é efetuado o descarte de atributos irrelevantes até que o conjunto de melhor desempenho seja encontrado. Por não trabalhar de maneira recursiva, o método *embedded* é menos custoso

que o método *wrapper*.

Figura 4 – Visão geral do funcionamento do método *Embedded*



Fonte: Adaptado de (DASH; LIU, 1997)

4.1.3 Trabalhos relacionados

Técnicas de seleção de atributos têm sido utilizadas com sucesso para melhorar o desempenho de modelos preditivos em problemas na área médica (REMESEIRO; BOLON-CANEDO, 2019). Alguns trabalhos no contexto de predição de mortalidade de pacientes que utilizam essas técnicas são descritos a seguir.

No trabalho de (POURHOMAYOUN; SHAKIBI, 2021), os autores empregam algoritmos supervisionados na fase de triagem por COVID-19 para prever o risco de mortalidade de pacientes. O conjunto de dados inicial dispõe de um total de 112 atributos. A principal contribuição apresentada pelos autores é um processo de seleção de atributos mais relevantes baseado no método *Filter*, no qual foram empregados como testes estatísticos para fins de avaliação, o coeficiente de correlação de Pearson, entropia e chi-square. O resultado do processo selecionou 57 dos 112 atributos originais, destacando a hipertensão e a idade como os atributos mais relevantes, ou seja, com maior valor nas análises estatísticas. A partir da seleção dos melhores atributos, O melhor desempenho dentre os algoritmos de classificação examinados foi alcançado pelo classificador Random Forest, o qual obteve uma AUC de 0.94 e uma probabilidade de classificação positiva de até 88%. O contexto do classificador apresentado neste trabalho é limitado à triagem de pacientes.

O trabalho apresentado por Lee (2018) testou a hipótese de que redes neurais profundas treinadas com atributos podem prever a possibilidade de mortalidade hospitalar pós-operatória. Os dados usados para treinar e validar o modelo de predição foram coletados de 59.985 pacientes, com 87 atributos extraídos ao final da cirurgia. Os autores apontaram como limitações a este estudo o tamanho da amostra, considerada um tanto limitada para o domínio de aplicação, quando comparado com modelos treinados com milhões de exemplos no dataset. Para resolver essa limitação e evitar *overfitting*, devido a um desbalanceamento de classes, as técnicas de seleção de atributos utilizadas foram as de regularização, um método do tipo *embedded* comumente

utilizado em *deep learning*. O resultado do processo de seleção de atributos culminou com a indicação de 45 atributos mais relevantes e um incremento da AUC do modelo de 0.81 para 0.91.

Mustafa et al., (2021) tiveram como objetivo principal criar um modelo de classificação para facilitar a identificação de quais pacientes internados com suspeita de COVID-19 estão em estado grave e precisam de serviços de saúde prioritários. O modelo utilizou uma estratégia constituída de duas etapas, em que na primeira um método *filter* foi aplicado para ordenar um conjunto de 120 atributos de entrada de acordo com sua relevância, por meio do cálculo de de correlação de Pearson reduzindo o conjunto para 58 atributos. Em seguida, foi aplicado um método *wrapper* com classificador de Árvore de Decisão que resultou em um subconjunto de 30 atributos e uma AUC = 0.85. .

A partir dos experimentos realizados pelos autores, pode-se destacar alguns pontos importantes. O principal deles é como modelos de classificação se beneficiam da utilização de técnicas de seleção de atributos. Também é importante destacar que existe uma diversidade de cenários específicos que lidam com predições de risco de complicação de saúde ou mortalidade de pacientes, ressaltando a necessidade de experimentos diferenciados, de acordo com o domínio do problema.

4.2 Metodologia

Esta seção descreve a metodologia empregada na análise da importância dos atributos selecionados como mais relvantes e descritos no Capítulo 3, dentro da abordagem de predição de risco de óbito de pacientes idosos com PAC.

4.2.1 Contexto do estudo e escopo do problema

Este estudo objetiva identificar, por meio de técnicas de seleção de atributos, aqueles que mais impactam na otimização do desempenho do modelo preditivo apresentado no Capítulo 3, especialmente no tocante à métrica AUC, tendo em vista esta ser a medida principal para avaliação do desempenho do classificador. Para fins de comparação, utiliza-se o valor de AUC igual a 0.81 obtido pelo classificador LR como *baseline*. Este estudo apresenta novas avaliações experimentais centradas na inclusão das técnicas de seleção de atributos descritas na Seção 4.1.2.

4.2.2 Cenários para experimentação

Com base no conjunto inicial de 29 atributos (Apêndice B, Tabela 1), os cenários definidos para experimentação têm como objetivo principal identificar os atributos mais relevantes que otimizam a precisão do modelo preditivo obtido neste trabalho. Para isso, a avaliação experimental busca comparar os três métodos de seleção de atributos (*Filter*, *Wrapper* e *Embedded*) aplicados conjuntamente ao treinamento do modelo e verificar o comportamento deste para

cada um dos métodos. Foram definidos cinco cenários para experimentação, conforme descrição seguinte:

Cenário 1 - relacionamento entre atributos e a classe do problema: Com o objetivo de obter os primeiros indícios acerca da correlação entres os atributos, esse cenário se baseia no cálculo da correlação de Pearson entre atributos. Nesse método foram avaliados os níveis de importância dos atributos em uma escala de 0 a 1, sendo 0 caracterizada como sem impacto para discriminar a classe do problema e 1 como determinante para a predição.

Cenário 2 - Filter: Este cenário tem a finalidade de estabelecer um ranqueamento inicial dos atributos mais discriminativos para designar a classe positiva do problema de classificação, antes mesmo de seu treinamento. O conjunto de dados foi avaliado utilizando a biblioteca *Python Sklearn*, com a função *SelectKBest*. A métrica utilizada para mensurar o nível de importância dos atributos nesse cenário foi a Analysis Of Variance (ANOVA) (função *f_classif*), recomendada para modelos de classificação com atributos numéricos e categóricos. O parâmetro de entrada corresponde aos top **N** atributos retornados pelo método. Para esse cenário foram experimentados os valores N = 5, N = 10 e N = 15, valores selecionados partindo do valor de correlação maior dos 10 atributos mais fortes identificados no Cenário 1, com uma margem de variação de 5 para melhor análise dos resultados.

Cenário 3 - Wrapper: O terceiro cenário objetiva realizar a seleção de atributos utilizando o método *wrapper* associado a um algoritmo de classificação. Neste cenário, foi escolhido o uso do algoritmo RFE com o intuito de selecionar e ranquear o número desejado de atributos para predição da classe alvo conforme o modelo de classificação utilizado, a saber: SVM, RF, MPL e LR.

Cenário 4 - Embedded: Este cenário teve como meta analisar a seleção de atributos com base no método *Embedded* aos modelos RF, LR e SVM. Neste experimento, em particular, justifica-se a execução com os modelos citados haja vista que possuem o método *feature_importances_* ou o método *coef_* embutidos, necessários para execução do experimento com a função da biblioteca *Sklearn*.

Cenário 5: Enquanto os cenários anteriores foram planejados com o objetivo de apontar o conjunto de atributos mais relevantes sob a perspectiva de seus métodos de seleção de atributos, o quinto experimento teve a finalidade de executar uma nova rodada de treinamento e teste do modelo *Logistic Regression* com os top **n** atributos mais relevantes determinados nos experimentos relacionados aos cenários 2, 3 e 4. A avaliação foi realizada por meio da métrica AUC para n=5, n=10 e n=15, e análise do gráfico ROC. O resultado foi comparado com o *baseline* (AUC igual a 0.81) obtido no experimento inicial descrito no Capítulo 3.

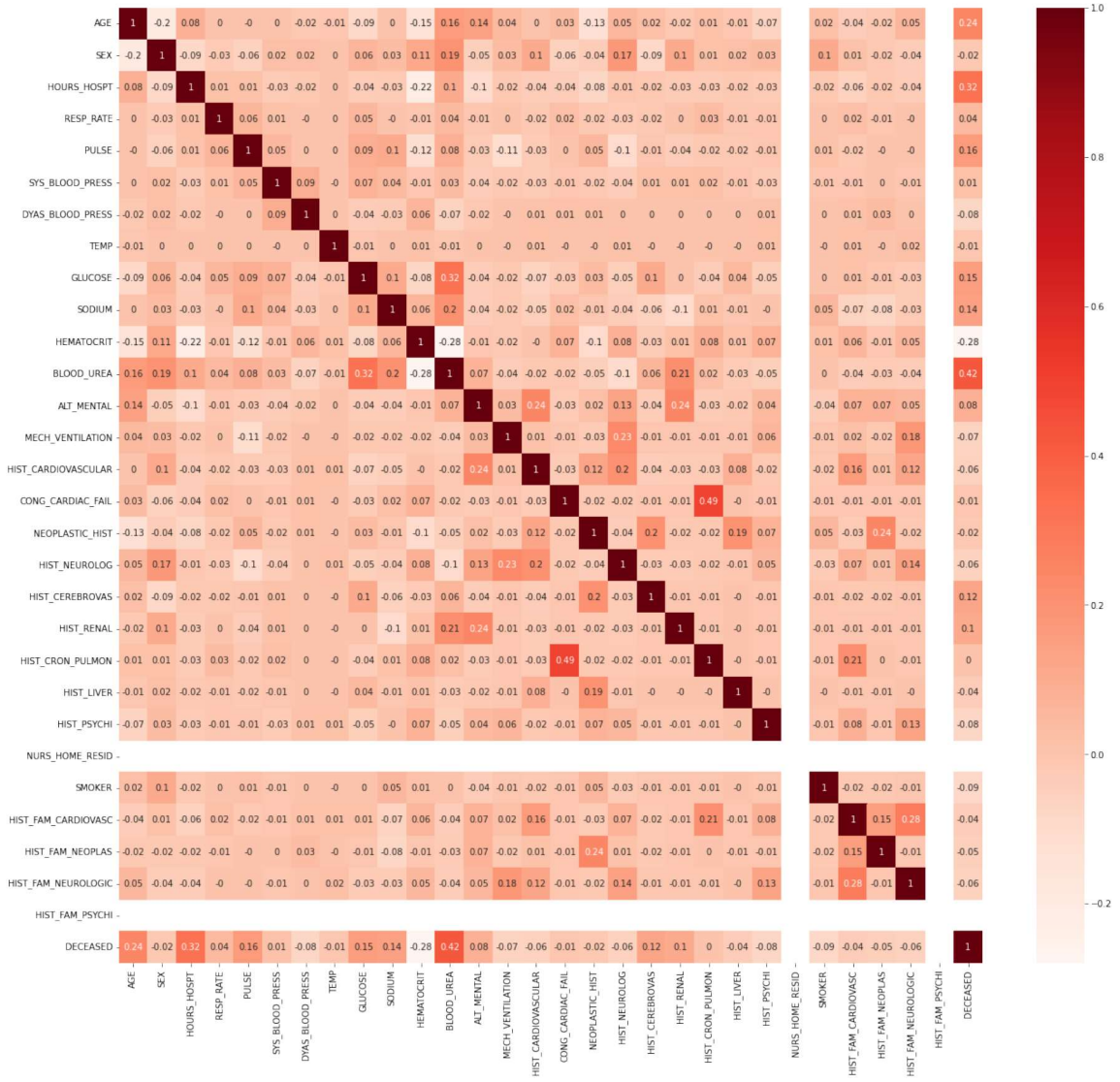
4.3 Resultados

Os resultados obtidos na avaliação experimental para os cinco cenários são mostrados a seguir.

4.3.1 Cenário 1: Correlação de Pearson

A Figura 5 mostra um mapa de calor com os valores do cálculo da correlação de Pearson entre pares de atributos, variando no intervalo real entre -1, 0 e 1.

Figura 5 – Mapa de cálculo da Correlação de Pearson entre atributos



Com relação à classe alvo do modelo preditivo, denominada na Figura 5 como "DECEASED", percebe-se que o maior valor positivo de correlação encontrado é de 0.42 atribuído ao atributo nível de uréia no sangue (BLOOD_UREA) do paciente. Outros atributos que se destacam em relação à classe alvo são o de tempo de hospitalização (HOURS_HOSPT) com correlação de 0.32 e o percentual de hematócritos no sangue (HEMATOCRIT) do paciente

com 0.28. Isto quer dizer que um aumento na medição das variáveis citadas implica em uma tendência de exercer um impacto de fraco a moderado para a determinação da classe positiva. Constata-se, também, uma relação linear no valor 0.49 entre os atributos "histórico de doenças crônicas pulmonares" (HIST_CRON_PULMON) e de "doenças cardíacas nos pacientes" (HIST_CARDIOVASCULAR). Os níveis de glicose (GLUCOSE) e de uréia no sangue (BLOOD_UREA) também mostram uma correlação de fator 0.42. Uma correlação negativa de -0.28 (BLOOD_UREA e HEMATOCRIT) e -0.22 (HEMATOCRIT e HOURS_HOSPT) também sugere que, respectivamente, o percentual de hematócritos no sangue em relação ao nível de uréia no sangue e o tempo de hospitalização, são inversamente proporcionais e determinam uma correlação fraca. Considerando a matriz de correlação apresentada, é possível inferir os primeiros indícios de atributos mais relevantes. Nenhum atributo obteve uma correlação positiva maior do que 0.5, observa-se a necessidade de um estudo mais abrangente de atributos para identificar outros indicativos no contexto de mortalidade de pacientes com pneumonia.

4.3.2 Cenário 2: Método filter com SelectKbest(F_classif)

A execução do método *Filter* aplicado com a função *F_classif* a partir dos 29 atributos de entrada. A partir da observação dos 10 atributos que se destacaram com valor de correlação com a classe alvo acima de 0.1 no cenário 1, foram selecionados os valores de N = 5, N = 10 e N = 15 como parâmetros para os N atributos mais relevantes. Os resultados são apresentados na Tabela 4.

Tabela 4 – N-Atributos selecionados pelo método *SelectKbest*

#	**Atributos**		
	N = 5	**N = 10**	**N = 15**
1	AGE	AGE	AGE
2	HOURS_HOSPT	HOURS_HOSPT	HOURS_HOSPT
3	PULSE	PULSE	PULSE
4	HEMATOCRIT	GLUCOSE	DYAS_BLOOD_PRESS
5	BLOOD_UREA	SODIUM	GLUCOSE
6		HEMATOCRIT	SODIUM
7		BLOOD_UREA	HEMATOCRIT
8		DYAS_BLOOD_PRESS	BLOOD_UREA
9		ALT_MENTAL	ALT_MENTAL
10		MECH_VENTILATION	MECH_VENTILATION
11			HIST_CEREBROVAS
12			HIST_RENAL
13			HIST_PSYCHI
14			SMOKER
15			HIST_FAM_NEUROLOGIC

Percebe-se que os atributos mais relevantes para cada valor de N se assemelham com as

percepções observadas no Cenário 1, haja vista a sobreposição de atributos em comum destacados na ordem decrescente pela análise de correlação. Ainda, conforme Tabela 4, o resultado mostra que o atributo "idade do paciente"(AGE) é o atributo melhor ranqueado na ordenação, seguida do tempo que o paciente está hospitalizado (HOURS_HOSPT), o nível de ureia no sangue (BLOOD_UREA) e o pulso (PULSE), para os três valores de N avaliados.

4.3.3 Cenário 3: Método Wrapper com RFE

Para este cenário de experimentação, foi executado o método RFE com os algoritmos MPL, SVM, LR e RF. O parâmetro selecionado para o método RFE foi definido com o valor de N = 10 atributos mais relevantes a serem apresentados, considerando os resultados obtidos nos Cenários 1 e 2. Devido ao fato do método *wrapper* ser mais custoso computacionalmente para realizar a seleção dos n atributos mais relevantes a partir da aplicação do algoritmo RFE, não foi possível realizar, em tempo hábil, o processamento de seleção de atributos para todos os valores de N. O resultado com os atributos selecionados pelo método *Wrapper* por algoritmo de classificação está retratado na Tabela 5.

Tabela 5 – Atributos selecionados pelo método Wrapper com o algoritmo RFE

#	**Atributos**			
	MLP	**SVM**	**LR**	**RF**
1	HOURS_HOSPT	BLOOD_UREA	AGE	HOURS_HOSPT
2	BLOOD_UREA	HOURS_HOSPT	HOURS_HOSPT	AGE
3	GLUCOSE	AGE	PULSE	PULSE
4	AGE	GLUCOSE	GLUCOSE	SODIUM
5	PULSE	SODIUM	SODIUM	GLUCOSE
6	HEMATOCRIT	HEMATOCRIT	HEMATOCRIT	HEMATOCRIT
7	SODIUM	PULSE	BLOOD_UREA	BLOOD_UREA
8	ALT_MENTAL	ALT_MENTAL	ALT_MENTAL	DYAS_BLOOD_PRESS
9	MECH_VENTILATION	DYAS_BLOOD_PRESS	MECH_VENTILATION	ALT_MENTAL
10	DYAS_BLOOD_PRESS	MECH_VENTILATION	DYAS_BLOOD_PRESS	MECH_VENTILATION

Pode-se observar que a relação de atributos resultantes é similar ao que foi apresentado no experimento referente ao Cenário 2, diferenciando em questão de ordem de classificação dos atributos. Além disso, o resultado está alinhado com as primeiras análises realizadas pela interpretação do mapa de calor do Cenário 1. Os atributos referentes a glicose (GLUCOSE), tempo de hospitalização (HOURS_HOSPT), idade (AGE), pulso (PULSE), hematócritos (HEMATOCRIT, sódio (SODIUM) e uréia no sangue (BLOOD_UREA) são comuns nos resultados dos Cenários 1 e 2.

4.3.4 Cenário 4: Métodos Embutidos

Para execução dos experimentos de análise de atributos com métodos embutidos, apenas os algoritmos RF e SVM possuíam as funções necessárias para sua execução. A relação ordenada dos atributos conforme sua importância é mostrada na Tabela 6.

Tabela 6 – Atributos selecionados pelo método *Embedded*

#	**Atributos**	
	SVM	**RF**
1	AGE	HOURS_HOSPT
2	HOURS_HOSPT	GLUCOSE
3	PULSE	ALT_MENTAL
4	DYAS_BLOOD_PRESS	AGE
5	GLUCOSE	MECH_VENTILATION
6	SODIUM	BLOOD_UREA
7	HEMATOCRIT	DYAS_BLOOD_PRESS
8	BLOOD_UREA	HIST_FAM_NEOPLAS

A quantidade de atributos a ser retornada não foi passada como parâmetro, utilizando assim o valor base pré-definido de 8 atributos no método executado. Nos atributos retornados como relevantes, percebe-se que tempo hospitalizado (HOURS_HOSPT), pulso (PULSE), glicose (GLUCOSE), sódio (SODIUM), hematócrito (HEMATOCRIT) e ureia no sangue (BLOOD_UREA) tornaram-se evidentes, assim como nos resultados dos cenários de experimentação anteriores. A ordem de importância desses atributos se modifica devido à aleatoriedade da divisão do conjunto de dados em treinamento e teste e diferentes métodos de execução.

4.3.5 Cenário 5: Avaliação do modelo preditivo com atributos selecionados

De modo a selecionar o conjunto ideal de atributos para um novo treinamento do classificador LR, a princípio foi estabelecida uma frequência de distribuição da presença dos atributos ao longo dos cenários de experimentação, como pode ser visto na Tabela 7

Pode ser observado na Tabela 7 que 13 dos 29 atributos originais foram selecionados por se encontrarem na interseção de todos os resultados dos cenários de experimentação. Com base nessa interseção, os dez atributos melhores ranqueados foram selecionados e utilizados para novo treinamento do classificador LR sob a justificativa de constarem mais de uma vez nos resultados dos experimentos relativos aos Cenários de 2 a 4. O desempenho do classificador, em termos de Curva ROC, treinado com o conjunto de dados reduzido aos 10 atributos mais relevantes, pode ser observado na Figura 6.

Como pode ser observado na Figura 6, o desempenho do melhor classificador (LR) com 10 atributos teve um incremento em relação à AUC de 0.05, passando de 0.81 (conjunto completo de 29 atributos) para 0.86.

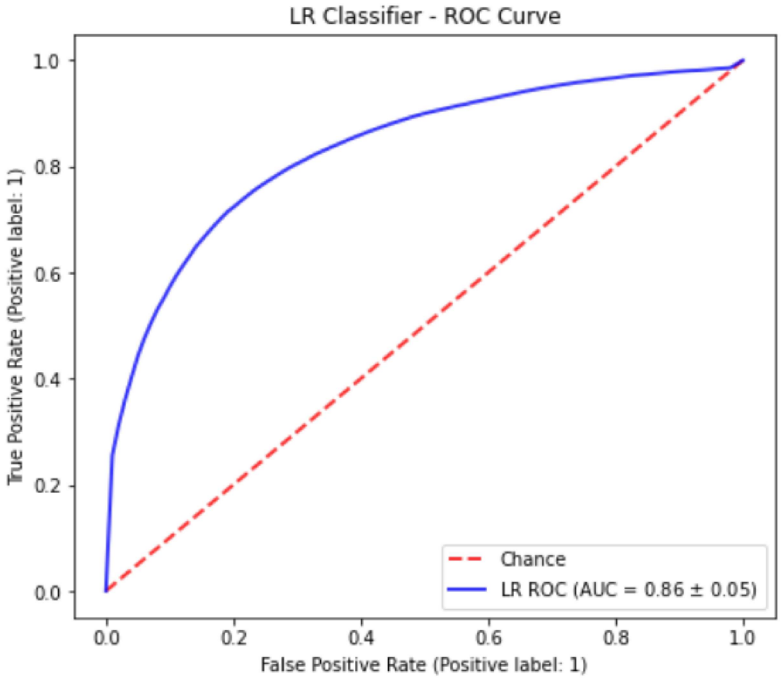
4.3.6 Análise dos cenários

Nos resultados dos experimentos obtidos na execução do treinamento e teste do modelo original, foi observado que os atributos tempo de hospitalização (HOURS_HOSPT), nível de

Tabela 7 – Frequência de distribuição dos atributos em todos os cenários de experimentos

#	Atributos	Quantidade
1	AGE	4
2	HOURS_HOSPT	4
3	PULSE	4
4	GLUCOSE	4
5	SODIUM	4
6	HEMATOCRIT	4
7	BLOOD_UREA	4
8	DYAS_BLOOD_PRESS	3
9	ALT_MENTAL	3
10	MECH_VENTILATION	3
11	HIST_CEREBROVAS	1
12	HIST_RENAL	1
13	SMOKER	1

Figura 6 – Curva ROC do classificador RL com 10 atributos



uréia no sangue e idade influenciaram para determinar a predição da classe positiva. Idade e tempo de internação são diretamente proporcionais à porcentagem de risco do paciente. Também foi observado que, em casos de pacientes com mais de 15 dias de internação, o modelo preditivo os classifica positivamente com percentuais de certeza acima de 70%. Os cenários de experimentação de seleção de atributos confirmaram os padrões de comportamento observados no modelo de classificação, visto que o tempo de internação e ureia no sangue formaram um grupo de interseção dos atributos selecionados em todos os métodos executados nos cenários descritos.

Destaca-se também que todos os métodos de seleção apresentam uma lista de atributos bem semelhantes, demonstrando uma consistência na participação desses atributos no conjunto de dados para treinamento e teste. Vale salientar que alguns atributos excluídos pelos métodos de seleção são muito relevantes no contexto de pneumonia, como é o caso do histórico de tabagismo (SMOKER), das doenças cardíacas (CONG_HEART_FAIL) ou mesmo a temperatura corporal (TEMP). A ausência de impacto desses atributos no modelo em si, indica uma inconsistência destes dados provavelmente obtida no período de extração ou até mesmo relacionado ao nível de detalhe da informação anotada no prontuário do paciente. Cabe uma análise do conjunto de dados original do hospital, juntamente com especialistas de domínio para inferir uma justificativa mais precisa sobre o impacto real que esses atributos poderiam exercer para o classificador. Isso poderá ser realizado em um trabalho futuro.

Em linhas gerais, os resultados dos experimentos indicaram o conjunto de atributos considerados mais relevantes para um melhor desempenho do classificador, superando o modelo treinado com o conjunto completo de atributos. O resultado aponta também a influência desses atributos em um possível diagnóstico assistido pelo classificador aos médicos. A análise realizada indica correlações entre atributos que, se aprofundadas por meio de um estudo com especialistas de domínio, podem indicar padrões de comportamento para os sinais vitais de pacientes. Como exemplo, a correlação negativa entre tempo de internação (HOURS_HOSPT) e hematócritos (HEMATOCRIT), se confirmada pela equipe médica, pode implicar em uma estratégia para manter o nível hematócrito no paciente por meio de medicação mais intensificada após uma quantidade específica de tempo.

5 CONSIDERAÇÕES E TRABALHOS FUTUROS

Prever o risco de óbito em pacientes idosos é um problema importante que exige atenção e cuidado, especialmente quando se trata de diagnósticos de pneumonia. Baseado nisso, foi criada uma abordagem focada em aprendizado de máquina para classificar pacientes em risco de óbito para PAC. Os resultados obtidos nesta dissertação incluem: (i) extração de dados e atributos específicos de anotações médicas e de enfermagem; (ii) definição de um baseline conforme um escore real (CURB-65) utilizado em hospitais; (iii) criação de uma abordagem de classificação que supera em termos de desempenho o baseline com relação à métrica AUC; (iv) realização de um teste de significância estatística para confirmar o melhor desempenho do classificador em comparação com o baseline definido; e (v) um estudo acerca da importância dos atributos de entrada do classificador.

Com respeito à primeira avaliação dos modelos de classificação, o classificador baseado em Regressão Logística foi capaz de prever o risco de mortalidade com o melhor desempenho de AUC=0.81. Este classificador também obteve média de probabilidade de confiança na predição da classe positiva de 78%. Os resultados do classificador foram superiores aos valores alcançados pelo baseline definido (AUC=0.61 e média de probabilidade de 20%). Os resultados também demonstraram que os pacientes internados com mais de 30 dias no hospital foram classificados com um risco de morte significativamente mais elevado, o que indica a importância de intensificação de cuidados médicos à medida que aumenta o tempo de internação e a própria contribuição do atributo para a criação do modelo de classificação. O teste estatístico realizado confirmou que a abordagem deste trabalho tem significância estatística superior ao baseline.

No estudo de importância de atributos, foram apresentados experimentos utilizando os métodos de seleção de atributos Filter, Wrapper e Embedded, nos quais foi possível estabelecer o conjunto final de atributos que consegue melhorar o desempenho do modelo preditivo. O desempenho do classificador, em termos de AUC, baseado em Regressão Logística foi aumentado para 0.86 com o subconjunto de dez atributos escolhidos que mais se destacaram em todos os cenários de experimentação.

Os resultados obtidos tanto na proposta do classificador quanto no estudo de atributos podem fornecer à equipe médica subsídios e um recurso poderoso capaz de indicar pacientes em risco de mortalidade por PAC com melhor nível de precisão em relação ao CURB-65, auxiliando na tomada de decisões e, como consequência, provendo um tratamento mais assertivo aos pacientes.

Como trabalhos futuros, sugere-se a realização de um estudo mais aprofundado para melhor entendimento do porque o conjunto com 10 atributos teve melhor desempenho e, principalmente, encontrar respostas para a exclusão de atributos considerados importantes na monitoração

do estado clínico do paciente terem tido menor importância do que o esperado. Também foram indicadas correlações entre atributos que podem ser estudadas com apoio de especialistas de domínio para conclusões mais precisas sobre as indicações estabelecidas. Outros atributos também podem ser estudados e incluídos no conjunto de dados para analisar o desempenho do modelo de classificação e, possivelmente, proporcionar resultados que podem ser ainda mais assertivos.

REFERÊNCIAS BIBLIOGRÁFICAS

AGGARWAL, D.; BALI, V.; MITTAL, S. An insight into machine learning techniques for predictive analysis and feature selection. In: . [S.l.: s.n.], 2019. v. 8, p. 342–349. Citado 2 vezes nas páginas 22 e 23.

BELLMAN, R. *Dynamic programming*. [S.l.]: Science, 1966. Citado na página 21.

DASH, M.; LIU, H. Feature selection for classification, intelligent data analysis. Elsevier, p. 131–156, 1997. Citado 3 vezes nas páginas 22, 23 e 24.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. The elements of statistical learning: Data mining, inference, and prediction. In: ____. Springer, 2017. Disponível em: <https://web.stanford.edu/~hastie/ElemStatLearn/printings/ESLII_print12_toc.pdf>. Citado na página 19.

HESPANHOL, V.; BáRBARA, C. Pneumonia mortality, comorbidities matter? *Pulmonology*, v. 26, n. 3, p. 123–129, 2020. ISSN 2531-0437. Disponível em: <https://www.sciencedirect.com/science/article/pii/S2531043719302053>. Citado na página 13.

LONG, B.; LONG, D.; KOYFMAN, A. Emergency medicine evaluation of community-acquired pneumonia: History, examination, imaging and laboratory assessment, and risk scores. *The Journal of Emergency Medicine*, v. 53, n. 5, p. 642–652, 2017. ISSN 0736-4679. Citado na página 13.

POURHOMAYOUN, M.; SHAKIBI, M. Predicting mortality risk in patients with covid-19 using machine learning to help medical decision-making. *Smart Health*, 2021. Citado na página 24.

PROVOST, F. Distributed data mining: scaling up and beyond, advances in distributed data mining. In: ____. [S.l.: s.n.], 2000. Citado na página 23.

REMESEIRO, B.; BOLON-CANEDO, V. A review of feature selection methods in medical applications. computers in biology and medicine. *Computers in biology and medicine*, v. 112, 2019. Citado 2 vezes nas páginas 22 e 24.

RYAN, L. et al. Mortality prediction model for the triage of covid-19, pneumonia, and mechanically ventilated icu patients: A retrospective study. In: *Anais de Medicine and Surgery*. [S.l.: s.n.], 2020. Citado na página 13.

SCHRÖER, C.; KRUSE, F.; GÓMEZ, J. M. A systematic literature review on applying crisp-dm process model. *Procedia Computer Science*, Elsevier, v. 181, p. 526–534, 2021. Citado na página 18.

SILVA, V. et al. Machine learning to assist in pneumonia decision making: A systematic review of the literature. In: SBC. *Anais do VIII Symposium on Knowledge Discovery, Mining and Learning*. [S.l.], 2020. p. 201–208. Citado na página 14.

SILVA, V.; SOUZA, D.; RÊGO, A. Predicting mortality risk among elderly inpatients with pneumonia: A machine learning approach. In: ICEIS. *Anais do 24° International Conference on Enterprise and Information Systems*. [S.l.], 2022. Citado na página 14.

VENKATESH, B.; ANURADHA, J. A review of feature selection and its methods. *Cybernetics and Information Technologies*, 2019. Citado 2 vezes nas páginas 21 e 22.

WIEMKEN, T.; KELLEY, R.; RAMIREZ, J. Clinical scoring tools: which is best to predict clinical response and long-term outcomes? *Infectious disease clinics of North America*, 2013. Citado na página 13.

World Health Organization. *Health Topics: Pneumonia*. 2015. Https://www.who.int/health-topics/pneumonia. Citado na página 13.

WU, D. et al. Risk factors of ventilator-associated pneumonia in critically iii patients. *Frontiers in pharmacology*, Frontiers, v. 10, p. 482, 2019. Citado na página 13.

Apêndices

APÊNDICE A – ARTIGO PUBLICADO NO KDMILE 2020

Machine Learning to Assist in Pneumonia Decision Making: A Systematic Review of the Literature

V. Monteiro Silva[1], A. Days Ramos Novo[1], D. Yluska de Souza[1], A. Sandro da Cunha Rêgo[2]

Programa de Pós Graduação em Tecnologia da Informação – IFPB - Campus JPA
João Pessoa – PB – Brazil
monteiro.victor@academico.ifpb.edu.br, days.amanda@academico.ifpb.edu.br,damires@ifpb.edu.br, alex@ifpb.edu.br

Abstract. Clinical decision support systems is a research area in which Machine Learning (ML) techniques can be applied. Nevertheless, specifically in assisting pneumonia decision making, the use of ML has not been so expressive. To help matters, this work aims to contribute to the evolution of the intersection of such areas by presenting a Systematic Review of the Literature. It provides results which may help to identify, interpret and evaluate how ML techniques have been applied and some research enhancements yet to be done.

CCS Concepts: • **Applied computing**;

Keywords: Data Mining, Machine Learning, Pneumonia, Systematic Review

1. INTRODUÇÃO

Pneumonia é uma categoria de infecção respiratória que pode atingir pessoas de todas as idades, especialmente crianças e idosos [Respira 2019]. Por ser uma das mais comuns infecções que geram necessidade de internação, casos dessa doença são considerados problemáticos, demandando uma atenção médica especial aos pacientes para evitar complicações, mortalidade e alta prematura durante o tramento do paciente [Rozenbauma et al. 2015].

A coleta de informações a partir do prontuário do paciente é uma rotina diária de profissionais de saúde para auxíliar no tratamento mais efetivo para pneumonia, assim como para qualquer outra enfermidade. O trabalho consiste em coletar e analisar fontes diversas de informação tais como sinais vitais, históricos familiares, anotações médicas, imagens de raio-x, dentre outras. Nesse contexto, uma parte significativa do treinamento médico é dedicada a aprender como identificar as informações relevantes desse montante, de modo que elas possam guiar a decisão sobre o tratamento [Bezemer et al. 2019]. Sistemas computacionais de suporte à decisão clínica (SAD) vêm sendo amplamente utilizados para auxiliar os profissionais de saúde nessa análise [Horng et al. 2017] [Porat et al. 2016].

A extração de conhecimento a partir do acervo de dados de pacientes pode trazer informações e padrões úteis de modo a agregar valor à tomada de decisão clínica no tratamento de pneumonia. Tendo em vista a inerente complexidade e diversidade em relação a dados e protocolos associados a tratamentos de pneumonia, a aplicação de técnicas de Mineração de Dados (MD) e de Aprendizado de Máquina (AM) junto aos SAD têm sido, cada vez mais, investigadas e utilizadas. É nesse cenário que ocorre esta pesquisa. A presente Revisão Sistemática da Literatura (RSL) objetiva apresentar um panorama acerca do estado da arte da aplicação de MD e AM com vistas à assistência à tomada de decisões em tratamentos de pneumonia. Busca-se assim mapear as evidências metodológicas que

concernem à convergência dessas áreas e prover a identificação de oportunidades de aprimoramentos de pesquisas nesse tema.

Este artigo está estruturado da seguinte forma: Na seção 2 são introduzidos alguns conceitos utilizados neste trabalho; na Seção 3 são discutidos os trabalhos relacionados; na Seção 4 é apresentada a metodologia utilizada na condução da RSL; Na Seção 5 é feita uma análise dos resultados obtidos e, na Seção 6, são explanadas as considerações finais acerca do trabalho.

2. FUNDAMENTAÇÃO TEÓRICA

Os Registros Médicos Eletrônicos (RME) são informações coletadas e armazenadas diariamente por diversos sistemas hospitalares. Eles incluem resultados de testes, diagnósticos, anotações médicas e medições de sinais vitais. A organização desses dados varia, podendo assumir valores numéricos, imagens de raios-x, textos livres de anotações médicas, entre outros [Ford et al. 2016]. Os dados utilizados nos RMEs podem ser classificados como dados estruturados (DE), não estruturados (DNE) e semi-estruturados (DSE). DE são dados que possuem esquemas de metadados bem definidos, como é o caso de bancos de dados relacionais [Santana 2019]. DNE não apresentam uma estrutura rígida devido à sua dinamicidade e DSE possuem uma estrutura intermediária entre os dois tipos anteriores. Parte considerável do trabalho dos profissionais de saúde é encontrar informação significativa nessa grande quantidade de dados para guiar os tratamentos.

Para facilitar a identificação de padrões, o processo *Knowledge Discovery in Database*s (KDD) dispõe da etapa de MD, que utiliza-se de técnicas computacionais para encontrar correlações, anomalias e padrões em grandes conjuntos de dados. O KDD sistematiza um conjunto de etapas que visa a descoberta de conhecimento potencialmente útil e previamente desconhecido a partir de um conjunto de dados, exibindo o conhecimento de maneira intuitiva. Além da etapa de MD, outras etapas norteiam o processo com atividades de seleção, pré-processamento, transformação, interpretação e avaliação dos resultados [Fayyad et al. 1996].

A MD é fundamentada tecnicamente no aprendizado de máquina (AM), subárea da Inteligência Artificial (IA) que lida com métodos computacionais e estatísticos para adquirir conhecimento de maneira automática [Mitchell 1997]. O AM é realizado a partir de um conjunto de dados cuidadosamente preparado e pode ser efetivado de modo supervisionado (AS) ou não supervisionado (ANS) [Monard and Baranauskas 2003]. O AS é baseado em um conjunto de observações para os quais as saídas desejadas são conhecidas ou em algum outro tipo de informação que represente o comportamento que deve ser apresentado ao sistema [de Souza Gomes 2019]. Exemplos de algoritmos comumente utilizados em AS incluem Redes Neurais Artificiais (RNA), *Support Vector Machine* (SVM), árvores de decisão, *Naive Bayes*, *K-Nearest Neighbors*, entre outros. O ANS, por sua vez, é baseado apenas nas observações do conjunto de dados de entrada cujos rótulos são desconhecidos, produzindo modelos que geralmente se destinam a explorar, agrupar ou identificar relacionamentos em comum entre suas entradas [Monard and Baranauskas 2003]. São exemplos de algoritmos aplicados no contexto de ANS: *K-means*, *K-medoids* e *Apriori*. Uma categoria do AM que vem sendo também utilizada na área de Saúde é a chamada Aprendizado Profundo (AP). O AP possui múltiplas camadas de processamento para extração de características e transformação de modelos de aprendizado. Seus algoritmos são mais complexos e possuem a capacidade de abstrair modelos de aprendizado mais representativos. São exemplos de algoritmos de AP: *Recurrent Deep Neural Networks* e *Convolutional Feedfoward Deep Neural Networks*.

3. TRABALHOS RELACIONADOS

A busca por trabalhos relacionados a esta RSL objetivou identificar trabalhos secundários associados ao uso de AM ou MD no auxílio à tomada de decisão em quadros de pneumonia. Alguns destes são brevemente descritos a seguir.

[Khan et al. 2020] apresentam uma RSL sobre a identificação de pneumonia a partir de imagens de raios-x utilizando AP. Como resultado, os autores evidenciam uma listagem das principais técnicas de deep learning encontradas e uma análise da qualidade e usabilidade de cada uma delas. Também são discutidos os datasets disponíveis e formas de balanceamento de dados utilizados.

O trabalho de [Chumbita et al. 2020] apresenta um estudo sobre o uso da IA para apoiar decisões clínicas no diagnóstico de pneumonia, utilizando imagens de raio-x. Observou-se que a maioria dos trabalhos empregou redes neurais com altas taxas de precisão nas previsões.

Uma abordagem baseada em modelos de classificação para prever readmissão em 30 dias após a alta hospitalar foi proposta por [Ben-Assuli and Padman 2018]. Os modelos foram treinados com dados dos pacientes e analisados sob a perspectiva de gráficos ROC (*Receiver Operating Characteristic*), tendo o modelo de Árvore de Decisão se destacado dentre os algoritmos analisados.

[Naydenova et al. 2016] apresentam uma abordagem de MD para previsão de pneumonia infantil utilizando de sinais vitais quantificáveis como atributos. Os experimentos demonstraram que o algoritmo Random Forests obteve melhor destaque na predição com métrica da AUC (*Area Under ROC Curve*).

Até onde foi possível identificar, não foram encontradas RSLs ou outros trabalhos secundários que apresentam panoramas da área de AM e MD aplicadas à assistência em quadros de pneumonia, com foco na análise de dados estruturados ou semi-estruturados.

4. METODOLOGIA

Essa RSL foi desenvolvida obedecendo às orientações, políticas e procedimentos estabelecidos por [Kitchenham 2004] e [Dybå and Dingsøyr 2008]. O processo é composto pelas seguintes etapas: (i) planejamento da RSL; (ii) condução da RSL, e (iii) resultados da RSL. Nesta seção, são mostrados os passos seguidos nas etapas (i) e (ii).

4.1 Planejamento da Revisão

A principal questão de pesquisa (QP) a ser tratada neste trabalho é: como a MD e o AM estão sendo aplicados no apoio à tomada de decisão clínica em quadros de pneumonia? Essa questão foi decomposta em questões específicas de pesquisa, conforme descrição seguinte:

QP1: Quais categorias de assistência à tomada de decisão em quadros de pneumonia são alvo das pesquisas?

QP2: Quais tarefas de MD foram identificadas para as categorias de assistência?

QP3: Quais modelos de AM encontrados para cada tarefa?

QP4: Quais métricas utilizadas para avaliação dos modelos?

Utilizando palavras-chaves relacionadas às questões de pesquisa, a estratégia de busca foi definida a partir de uma string genérica: ("machine learning" OR "data mining" OR "deep learning") AND ("pneumonia"). As bases de dados escolhidas para as pesquisas foram: ACM Digital Library, IEEE, Springer, PubMed, Science Direct e CAPES. As fontes escolhidas estão entre as principais bases relacionadas à QP que disponibilizaram listagens indexadas dos artigos para análise.

Os critérios de inclusão (CI) usados como filtros foram: (i) estudos que apresentem a aplicação de técnicas de MD em tratamento ou prevenção de pneumonia e (ii) estudos que respondem pelo menos uma questão específica de pesquisa. Os critérios de exclusão (CE) utilizados foram: (i) estudos que não provêem relevância científica; (ii) artigos publicados antes de 2010; (iii) estudos secundários ou terciários; (iv) estudos sem experimentação ou avaliação e (v) estudos que não consideram dados estruturados ou semi-estruturados.

Foram adicionalmente definidos critérios de qualidade (CQ) a serem aplicados aos trabalhos selecionados, a saber: (i) definição clara dos objetivos do estudo; (ii) relevância em relação à utilização de técnicas de ML para suporte à decisão clínica; (iii) utilização de métodos de coleta de dados; (iv) apresentação de resultado coerentes com os objetivos e (v) fundamentação teórica sobre os tópicos do estudo. Para cada CQ foi atribuída uma escala de valores: 0, quando o critério é ausente ou não aplicado; 1, quando é parcialmente atendido e 2, quando é totalmente atendido.

4.2 Condução da Revisão

A condução da revisão, ou seja, a execução do protocolo planejado, ocorreu em dois meses. A seleção dos trabalhos foi realizada em três fases. Iniciada a partir da seleção de estudos pela string geral de busca, a fase 1 trouxe 563 resultados. Em seguida foi realizado a leitura do título e do resumo dos artigos e, foram excluídas as ocorrências que não respondiam pelo menos uma questão de pesquisa. Na fase 2, foi realizada a leitura das seções de introdução e conclusão dos artigos, resultando em 241 trabalhos selecionados. Esta fase também eliminou os estudos que não estavam alinhados com os objetivos desta RSL. Por fim, os artigos submetidos à terceira fase foram integralmente lidos e analisados, resultando em 34 estudos. Esses trabalhos foram avaliados, quantificados com uma nota conforme os critérios de qualidade definidos e respectivos valores e utilizados para a extração das informações esperadas e definidas no protocolo.

5. RESULTADOS

Os 34 artigos selecionados foram avaliados conforme os CQ. Desses, 19 trabalhos atingiram a nota máxima com 10 pontos. A Tabela I mostra os quantitativos dos trabalhos selecionados por fase, separados por fonte de busca. Os resultados particularizados por Questão de Pesquisa são descritos logo adiante.

Table I. Análise Quantitativa.

Fonte	**Fase 1**	**Fase 2**	**Fase 3**
ACM Digital Library	8	8	2
IEEE Xplore	42	29	5
Springer	59	55	15
PubMed	250	70	2
Science Direct	92	33	8
CAPES	112	46	2
TOTAL	**563**	**241**	**34**

5.1 Quais categorias de assistência à tomada de decisão em quadros de pneumonia são alvo das pesquisas?

As categorias de acordo com as principais contribuições identificadas como resultado do planejamento da RSL foram:

—**C1.Detecção e Diagnóstico**. Trabalhos com foco em detecção e diagnóstico de pneumonia se concentram na aplicação de técnicas de AM que utilizam dados clínicos dos pacientes para prever a ocorrência da doença;

—**C2.Mortalidade e Complicações**. Essa categoria se refere aos trabalhos que tinham como objetivo sinalizar os pacientes indicativos de risco de falecimento ou de complicações que possam conduzi-los à admissão na UTI ou à necessidade de ventilação mecânica;

—**C3.Atributos e Biomarcadores**. Referente a trabalhos exploram a análise de atributos e biomarcadores, com a finalidade de identificar dados clínicos que têm forte correlação com o diagnóstico, tratamento ou complicações do quadro dos pacientes com pneumonia;

—**C4.Tempo de Internação e Readmissão**. Trabalhos enquadrados nesta categoria têm como objetivo avaliar o tempo de internação necessário do paciente ou prever um possível risco de readmissão hospitalar.

Os dados apresentados na tabela II expõem o quantitativo das categorias de assistência mais recorrentes .

Table II. Quantitativo de trabalhos por Categoria.

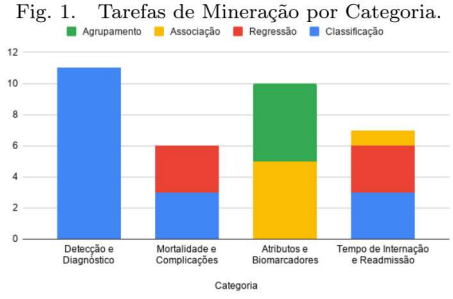
Categoria	**Quantitativo**
Detecção e Diagnóstico	11
Mortalidade e Complicações	6
Atributos e Biomarcadores	10
Tempo de Internação e Readmissão	7
TOTAL	**34**

Pode-se observar que a maioria dos trabalhos selecionados (32.4%) tem como alvo a realização de estudos acerca da Detecção e Diagnóstico de Pneumonia ou de patologias relacionadas. Categorias com menor representatividade são relacionadas a problemas de tempo de hospitalização e risco de readmissão, juntamente com a identificação de mortalidade e complicações nos pacientes já diagnósticados. Uma lacuna encontrada no estado da arte foi na estimativa numérica para o tempo de internação, aonde dentre diversas condições de pacientes, requerem diferentes estudos específicos para melhor assertividade. Ainda, na categoria C1, destacam-se em número os trabalhos no diagnóstico em crianças e idosos, duas faixas etárias com maior incidência de casos de pneumonia. Exemplos de trabalhos que se destacam com maior nota nos CQ neste quesito incluem [Naydenova et al. 2015], [DeLisle et al. 2013] e [Liao et al. 2020]. Outros trabalhos que se destacaram nas demais categorias com maior nota nos índices de qualidade foram:[Baechle et al. 2017], [Wu et al. 2014] e [Caruana et al. 2015].

5.2 Quais tarefas de MD foram identificadas?

O gráfico ilustrado na Figura 1 sumariza a parcela de ocorrência das tarefas de MD por categoria indicada na Subseção 5.1.

Fig. 1. Tarefas de Mineração por Categoria.



É perceptível na Figura 1 que os trabalhos com foco no diagóstico preventivo de pneumonia utilizaram apenas tarefas de classificação. Conforme comentado pelos autores, a identificação da patologia no início do tratamento é um fator essencial para um bom prognóstico e por isso a tarefa de classificação se destaca entre as demais [DeLisle et al. 2013] [Ge et al. 2019]. Para trabalhos focados nas

categorias de Mortalidade e Complicações e Tempo de Internação e Readmissão, existe uma divisão dos resultados entre Tarefas de Classificação, em que os autores focam apenas na sinalização desses riscos à equipe médica[Wu et al. 2014][Shimizu et al. 2019][Lai et al. 2018] e Regressão, para estimar a probabilidade de tais riscos ocorrerem [Caruana et al. 2015][Villiers et al. 2018]. No que se refere à identificação de Atributos ou Biomarcadores que influenciam no tratamento de pneumonia, autores utilizaram unicamente técnicas de Agrupamento e Regras de Associação, incluindo entre os dados analisados *scores* médicos e comparando ocorrências de atributos com Diagnósticos [Baechle et al. 2017] [Lin et al. 2010] [Ubaid et al. 2010].

5.3 Quais os modelos de AM encontrados para cada tarefa?

Identificadas as tarefas de MD mais recorrentes nos resultados, foram destacados os modelos de AM mais utilizados em cada uma das tarefas. De acordo com a Tabela III, a tarefa de classificação tem o destaque de utilização do algoritmo SVM na maioria dos trabalhos. Os algoritmos LR e RF também ficaram entre os três modelos mais utilizados. O consenso entre os trabalhos com maior nota nos CQ que utilizaram esses algoritmos é de que eles são comumente aplicados e na área de saúde e apresentam bom desempenho na tarefa de classificação, a exemplo do SVM em [Caruana et al. 2015][Lai et al. 2018][Wu et al. 2014]. Na tarefa de Regressão, os modelos RF, LR e AD se destacam dentre os demais modelos pelos mesmos motivos citados [Caruana et al. 2015] [Chmielewska 2016].

Table III. Utilização dos modelos por Tarefa.

Modelos	Classificação	Regressão	Associação	Agrupamento
SVM	12	1	0	0
RNA	5	1	0	0
RF	7	2	0	0
LR	7	2	0	0
AS	0	1	3	4
AD	3	2	0	1
NB	3	0	0	0
TOTAL	**37**	**9**	**3**	**5**
Legenda: SMV - Support Vector Machine; RNA - Rede Neural Artificial; RF - Random Forests; LR - Logistic Regression; AS - Análise Estatística; AD - Árvore de Decisão; NB - Naive Bayes;				

Nas Tarefas de Associação e Agrupamento, utilizadas para identificar atributos e biomarcadores, os modelos de análise estatísticas tiveram forte predominância, ressaltando lacunas no estado da arte aonde aplicação de algoritmos de AM focados nessas tarefas, poderiam apresentar resultados melhores. [Chen et al. 2016][Ubaid et al. 2010]. Outros modelos como *k-nearest neighbors*, algoritmo genético, Apriori, *k-means* e AP, de modo geral, também foram observados nos trabalhos, entretanto, todos tiveram uma representatividade baixa, com uma ou duas ocorrências, em comparação com os demais.

5.4 Quais as métricas mais utilizadas para avaliação dos modelos?

As principais métricas de avaliação dos modelos encontrados na análise dos resultados são destacadas na Tabela IV. Na maioria dos resultados, a plotagem da curva ROC foi o método gráfico mais empregado para avaliar o desempenho dos modelos de classificação.

Considerando que a análise da curva ROC é baseada na Sensitividade e Especificidade, representando a relação bidimensional entre os casos verdadeiros positivos e falso positivos , justifica-se sua utilização pela maioria dos modelos considerando que erros na previsão desses casos causam maior impacto os pacientes, representando doentes, corretamente classificados como doentes [Caruana et al. 2015] [Naydenova et al. 2015] [Lai et al. 2018] [Ge et al. 2019]. Acurácia geral e Score Kappa obtiveram a menor representatividade nos trabalhos, resultado justificável visto que ambas as medidas focam na

Table IV. Principais Métricas de avaliação por Modelo.

Modelos	ROC	ACC	SEN	ESP	KAPP
SVM	10	8	9	8	2
RNA	6	4	4	3	1
RF	9	5	6	6	2
LR	9	3	4	4	0
AS	0	0	0	0	0
AD	5	4	3	3	3
NB	3	2	2	2	2
TOTAL	**42**	**26**	**28**	**26**	**10**
Legenda: ROC - Curva ROC; ACC - Acurácia Geral; SEN - Sensitividade; ESP - Especificidade; KAPP - Score Kappa;					

performance genérica dos modelos, enquanto os modelos utilizados na área médica se concentram em errar o menos possível na classe de interesse.

6. CONSIDERAÇÕES FINAIS

Nesta pesquisa foi apresentado o planejamento, a condução e os resultados obtidos de uma RSL sobre a utilização de AM e MD para auxiliar na tomada de decisão clínica no tratamento de pneumonia. Ao final de cada etapa de seleção, foi realizada a revisão dos estudos aceitos garantindo assim a confiabilidade e reprodutibilidade da revisão. Foi feita uma categorização dos trabalhos encontrados, destacando diferentes aplicabilidades de pesquisas no tratamento de pneumonia. O trabalho também apresentou uma análise quantitativa destacando tarefas de MD e modelos de AM que foram mais utilizados, juntamente com as métricas de desempenho mais recorrentes para cada algoritmo. Observa-se, a partir dos trabalhos analisados e resultados, que a aplicação de AM e MD possui um alto potencial para fornecer conhecimento e valor importantes e úteis para assistência à tomada de decisão em quadros de pneumonia. A grande diversidade de condições clínicas de pacientes, cada um com cenários distintos de tratamento efetivo, mostra que essa é uma área a ser investigada e aprofundada. Uma lacuna encontrada inclui a falta de trabalhos para estivativa de tempo de internação recomendado. Verificou-se também que muitos dos trabalhos avaliados possuem datasets com dados de poucos pacientes, devido a falta de registros médicos eletrônicos ou a restrição pela sensibilidade da informação, o que foi destacado pelos autores como limitações nos seus estudos. As evidências obtidas neste trabalho servirão de norte para a evolução das pesquisas neste contexto.

REFERENCES

Baechle, C., Agarwal, A., and Zhu, X. Big data driven co-occurring evidence discovery in chronic obstructive pulmonary disease patients. *Journal of Big Data* vol. 4, pp. 9, 03, 2017.

Ben-Assuli, O. and Padman, R. Analysing repeated hospital readmissions using data mining techniques. *Health Systems* 7 (2): 120–134, 2018.

Bezemer, T., de Groot, M., Blasse, E., ten Berg, M., Kappen, T. H., Bredenoord, A. L., van Solinge, W. W., Hoefer, I. E., and Haitjema, S. A Human(e) Factor in Clinical Decision Support Systems. *Journal of Medical Internet Research* 21 (3): e11732, 2019.

Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., and Elhadad, N. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '15. Association for Computing Machinery, New York, NY, USA, pp. 1721–1730, 2015.

Chen, C., Shi, L., Li, Y., Wang, X., and Yang, S. Disease-specific dynamic biomarkers selected by integrating inflammatory mediators with clinical informatics in ards patients with severe pneumonia. *Cell Biol Toxicol*, 2016.

Chmielewska, M. Clostridium difficile infection due to pneumonia treatment: Mortality risk models. *in: Pokorski M. (eds) Pathobiology of Pulmonary Disorders*, 2016.

Chumbita, M., Cillóniz, C., Puerta-Alcalde, P., Moreno-García, E., Sanjuan, G., Garcia-Pouton, N., Soriano, A., Torres, A., and Garcia-Vidal, C. Can artificial intelligence improve the management of pneumonia. *Journal of Clinical Medicine* 9 (1): 248, Jan, 2020.

de Souza Gomes, E. A. Implementando um Mural Eletrônico em PHP: Uma Aplicação Voltada a uma Instituição de Ensino Superior. Aplicabilidade de Algoritmos de Aprendizado de Máquina para Detecção de Intrusão e Análise de Anomalias de Rede, 2019. Dissertação (especialização) — UFMG, Brasília, DF.

DeLisle, S., Bernard, K., Deepak, J., and Siddiqui, T. Using the electronic medical record to identify community-acquired pneumonia: toward a replicable automated strategy. *PLoS One* vol. 8, 08, 2013.

Dybå, T. and Dingsøyr, T. Empirical studies of agile software development: A systematic review. *Information and Software Technology* vol. 50, pp. 833–859, 08, 2008.

Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. From data mining to knowledge discovery in databases. *AI magazine* 17 (3): 37, 1996.

Ford, E., Carroll, J., Smith, H., Scott, D., and Cassell, J. Extracting information from the text of electronic medical records to improve case detection: A systematic review. *Journal of the American Medical Informatics Association* vol. 23, pp. ocv180, 02, 2016.

Ge, Y., Wang, Q., Wang, L., Wu, H., Peng, C., Wang, J., Xu, Y., Xiong, G., Zhang, Y., and Yi, Y. Predicting post-stroke pneumonia using deep neural network approaches. *International Journal of Medical Informatics* vol. 132, pp. 103986, 2019.

Horng, S., Sontag, D., Halpern, Y., Jernite, Y., Shapiro, N., and Nathanson, L. Creating an automated trigger for sepsis clinical decision support at emergency department triage using machine learning. Vol. 12, 2017.

Khan, W., Zaki, N., and Ali, L. Intelligent pneumonia identification from chest x-rays: A systematic literature review. *medRxiv* 12 (5): p–p, 2020.

Kitchenham, B. Procedures for Performing Systematic Reviews, 2004.

Lai, H., Chan, P., Lin, H., Chen, Y., Lin, C., and Hsu, J. A web-based decision support system for predicting readmission of pneumonia patients after discharge. In *2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. pp. 2305–2310, 2018.

Liao, Y.-H., Shih, C.-H., Abbod, M., and Shieh, J.-S. Development of an e-nose system using machine learning methods to predict ventilator-associated pneumonia. *Microsyst Technol 2020*, 03, 2020.

Lin, W.-T., Wang, S.-T., Chiang, T.-C., Shi, Y.-X., and yu Chen, W. Abnormal diagnosis of emergency department triage explored with data mining technology: An emergency department at a medical center in taiwan taken as an example. *Expert Systems with Applications* 37 (4): 2733 – 2741, 2010.

Mitchell, T. *Machine Learning*. McGraw-Hill International Editions. McGraw-Hill, 1997.

Monard, M. C. and Baranauskas, J. A. Conceitos sobre aprendizado de máquina. In *Sistemas Inteligentes Fundamentos e Aplicações*, 1 ed. Manole Ltda, Barueri-SP, pp. 89–114, 2003.

Naydenova, E., Tsanas, A., Casals-Pascual, C., and De Vos, M. Smart diagnostic algorithms for automated detection of childhood pneumonia in resource-constrained settings. In *2015 IEEE Global Humanitarian Technology Conference (GHTC)*. pp. 377–384, 2015.

Naydenova, E., Tsanas, A., Howie, S., Casals-Pascual, C., and de Vos, M. The power of data mining in diagnosis of childhood pneumonia. *Journal of The Royal Society Interface* vol. 13, pp. 20160266, 07, 2016.

Porat, T., Kostopoulou, O., Woolley, A., and Delaney, B. C. Eliciting user decision requirements for designing computerized diagnostic support for family physicians. Vol. 10, 2016.

Respira, A. A pneumonia. http://www.respira.pt/content/docs/pneumonia.pdf, 2019.

Rozenbauma, M., Mangenc, M.-J., Huijts, S., van der Werf, T., and Postma, M. J. Incidence, direct costs and duration of hospitalization of patients hospitalized with community acquired pneumonia: A nationwide retrospective claims database analysis. *Vaccine* 3 (28): 3193 – 3199, 2015.

Santana, B. S. *Extração e Aplicação de Indicadores no Processo de Recomendação de Recursos Urbanos Utilizando Dados Estruturados e Não-Estruturados*. M.S. thesis, Universidade Federal do Rio Grande do Sul, https://lume.ufrgs.br/handle/10183/193897, 2019.

Shimizu, S., Hara, S., and Fushimi, K. Prs55 predicting the risk of in-hospital mortality in adult community-acquired pneumonia patients with machine learning: A retrospective analysis of routinely collected health data. *Value in Health* vol. 22, pp. S882, 2019. ISPOR Europe 2019.

Ubaid, A., Mirza, F., Baig, M., and Mirza, F. Identifying the relationship between unstable vital signs and intensive care unit (icu) readmissions. *Expert Systems with Applications* 37 (4): 2733 – 2741, 2010.

Villiers, L., Caspar, Y., Marche, H., Boccoz, S., Maurin, M., Marche, P., Morand, P., Marquette, C., and Corgier, B. Resynplex: Respiratory syndrome linked pathogens multiplex detection and characterization. *IRBM* 39 (5): 368 – 375, 2018.

Wu, C., Rosenfeld, R., and Clermont, G. Using data-driven rules to predict mortality in severe community acquired pneumonia. *PLOS ONE* 9 (4): 1–9, 04, 2014.

APÊNDICE B – ARTIGO PUBLICADO NO ICEIS 2022

Predicting Mortality Risk among Elderly Inpatients with Pneumonia: A Machine Learning Approach

Victor Monteiro Silva, Damires Yluska de Souza Fernandes and Alex Sandro da Cunha Rêgo
Federal Institute of Paraíba, João Pessoa, Brazil
monteiro.victor@academico.ifpb.edu.br,{damires,alex}@ifpb.edu.br

Keywords: Data Analysis and Prediction, CAP, Probability of Death, ROC Curve, AUC.

Abstract: Community-acquired Pneumonia (CAP) is a serious respiratory infection that can cause life-threatening risk in people of different ages, especially in elderly inpatients. Regarding this age group, mortality rates by CAP still can reach 30% of all respiratory causes of death. In this work, we propose a machine learning approach to predict mortality risk among elderly inpatients with CAP. The approach uses real world data of elderly people with CAP from a hospital in Brazil, collected from 2018 to 2021. Based on patients data as learning features, our approach is able not only to classify patients at risk of mortality during hospitalization, but also to estimate the probability concerning the prediction. Some classification models have been examined and, among them, the best performance in terms of Area under ROC Curve (AUC) value has been achieved by the Logistic Regression (LR) classifier (AUC=0.81). Accomplished results show that the presented approach outperforms CURB-65 score as baseline in terms of both AUC values and probability of patient death. Besides, our approach is able to output probabilities ranging from 50 to 99% w.r.t. positive classification, i.e., patients that may come to death. A statistical test confirms that the presented approach outperforms the baseline provided by the CURB-65.

1 INTRODUCTION

Community-acquired Pneumonia (CAP) is a serious respiratory infection that can cause life-threatening risk in people of different ages (World Health Organization, 2015). As one of the most common infections that result in the need of hospitalization, it may inflame the air sacs in one or both lungs, affect other vital organs and cause difficult breathing. Cases of patients diagnosed with CAP are considered hard to deal with and it is more likely to have complications in this kind of disease if a patient is an older adult, a very young child, or if s/he has a weakened immune system, or a serious medical problem like diabetes or cirrhosis (Wu et al., 2019). Despite ever growing better health-care access, with not only medical science progress but also specialized units and sophisticated life-support systems, CAP mortality rates still can reach 30% of all respiratory causes of death mainly with regards to elderly inpatients (Hespanhol and Bárbara, 2020). Indeed it may be particularly severe in people ages 65 years or older, implying in a higher mortality risk when compared to other age groups.

Continuously analyzing patient data is a common task for health professionals to make decisions regarding treatments. However, identifying relevant information from the data is sometimes a challenging task (Bezemer et al., 2019). This process can also be time consuming, since it usually requires considering medical imaging exams, laboratory results, vital signs, patient history and also medical annotations. These data are often scattered throughout the hospital systems and databases (Wiemken et al., 2017).

To help matters in the decision making of health professionals, some medical scores have been defined and used. Regarding pneumonia treatments, two scores are commonly used, namely (Long et al., 2017): Pneumonia Severity Index (PSI) and CURB-65. Both scores provide a preliminary method for inpatient mortality prognoses, giving the medical team an alert based on Electronic Medical Record (EMR) data (Ryan et al., 2020). Nevertheless, these medical scores lack efficiency for individual patient-level decision making, since they only provide an estimate up to 27% of chance of patient mortality. This may be due to the fact that the score results only consider the current state of EMR data of a given patient, excluding his/her treatment evolution itself. Also it does not take into account other patient examples with similar

conditions in terms of general symptoms, signs, prognoses and progressions (Wiemken et al., 2013).

In this work, we aim to consider in what ways could a predictive analytical model help to address inpatient mortality risk problem in CAP cases. To this end, two aspects should be taken into account (Pourhomayoun and Shakibi, 2021): (i) the large and increasing volume of historical patient data, and (ii) the generation and usage of a model that generalises beyond the dataset in such a way that it may assist health professionals to make more assertive decisions on inpatients treatments. In this scenario, we define two main research problems that have guided our work: (i) How to identify elderly inpatients diagnosed with CAP at risk of death? And, in addition, (ii) how to provide the probability that such prediction may indeed occur?

In this sense, we propose a supervised learning approach to predict mortality risk with respect to elderly inpatients with CAP. Based on patients EMR data as learning features, our approach is able to classify patients at risk of mortality during hospitalization. In addition, it can estimate a probability of inpatients come to death, by means of a range from 50% to 99% w.r.t. positive classification (patients that do not survive). The approach uses real world data of elderly people with CAP from a hospital in Brazil, which were collected from 2018 to 2021 and prepared for usage in this work. We evaluate our approach under two aspects: (i) particularly analysing Receiver Operating Characteristic (ROC) curves, which are used in medicine to determine diagnostics effectiveness of classification models, and (ii) by computing ROC's Area Under the Curve (AUC), which provides the overall performance of the most critical classification in this work (patients classified as at risk of death). Accomplished results show that the presented approach outperforms CURB-65 score as baseline both in terms of AUC and of the obtained probability for risk of death. Results also bring to attention a time limit of hospitalization that hugely increased the probability of death, considering some chronological measurements of inpatients.

Our contributions are summarized as follows: (i) a relevant dataset built based on different factors correlated to pneumonia, including some features extracted from medical annotations; (ii) an approach using machine learning algorithms for analyzing and predicting risk of death in elderly inpatients with CAP; (iii) a baseline built based on a real medical score (CURB-65); (iv) a comparative evaluation between the computational version of a baseline and the best achieved classification model using ROC curves; (v) a statistical significance test, which confirms that our predictive model outperforms the baseline; and (vi) a data analysis w.r.t. a patient chronology regarding results achieved by the best classifier.

This paper is organized as follows: Section 2 provides some theoretical background; Section 3 describes some related works; Section 4 introduces aspects of the research methodology applied in this work; Section 5 presents the proposed approach with the experimental evaluation accomplished and results. Section 6 concludes the paper and points out some future work.

2 THEORETICAL BACKGROUND

CAP is a form of intense respiratory infection that affects the lungs. This can lead to symptoms such as cough and shortness of breath. In severe cases, hospitalization is rather recommended (World Health Organization, 2015)(Long et al., 2017). Particularly, there are some reasons why CAP can be more severe in older adults (World Health Organization, 2015): immune system naturally weakens as people age and older adults are more likely to have chronic health conditions, such as heart diseases, what can increase their risk for pneumonia. In order to improve patient care and management regarding CAP, medical professionals make use of inpatient risk scores.

A number of pneumonia severity scores have been described in the literature (Chen et al., 2010)(Long et al., 2017). Severity scores are important to ascertain, for instance, safety criteria to discharge/admit patients and time to remain in an Intensive Care Unit (ICU) (Webb and Gattinoni, 2016). These scores support clinical decision-making in a variety of scenarios and can be found in the literature to calculate the probability of morbidity and mortality among inpatients with pneumonia. The scores most commonly used are the CURB-65 and PSI (Long et al., 2017)(Chen et al., 2010). Both PSI and CURB-65 use data from patient medical records, such as laboratory results, vital signs and demographic data, in order to estimate mortality or even help determining inpatient versus outpatient treatment. To this end, they provide some categories of risk, based on the score calculation discussed in the following (Long et al., 2017)(Chen et al., 2010).

The CURB-65 scores range from 0 to 5 and includes points for each one of the following criteria, namely (Webb and Gattinoni, 2016): patient has confusion (defined by a mental test score); blood urea $>$ 20 mg/dL; respiratory rate $\geq$ 30 breaths/min; blood pressure (systolic $<$ 90 mm/Hg, or diastolic $\leq$ 60 mm/Hg) and age $\geq$ 65 years. Clinical management decisions can be made based on the resulting score,

which is achieved according to the following punctuation marks (Webb and Gattinoni, 2016):

- **1 point**: probably suitable for home treatment; low risk group: 2.7% mortality risk.
- **2 points**: consider hospital supervised treatment; Moderate risk group: 6.8% mortality risk.
- **3 points**: Consider ICU admission; Severe risk group: 14.0% mortality risk.
- **4 - 5 points**: Consider ICU admission; Highest risk group: 27.8% mortality risk.

The PSI medical score uses similar score points as CURB but it also includes additional features such as gasometry exam results. Management based on PSI is quite similar to CURB 65, although it provides some specific rules to ages above 50.

Although widely used and indeed useful, these scores only consider current EMR data of a given inpatient. It does not take into account other points such as the treatment evolution itself as well as examples of other similar cases and their prognoses (Wiemken et al., 2013). It may be rather important to consider not only the whole patient health history and his/her clinical stability, but also his/her individual risk factors for severe diseases, such as the case of pneumonia (Chen et al., 2010)(Long et al., 2017)(Wiemken et al., 2013).

Machine Learning (ML) provides computational and statistical methods to automatically acquire knowledge from data (Alpaydin, 2016). Solutions based on ML are developed from a carefully prepared dataset and commonly are performed by supervised or unsupervised learning methods. The value of ML in healthcare comes from its ability to process large amount of health care data to extract clinical insights that may be helpful to medical decision-making. Recent works exploring ML methods point out that predictive models have the potential for identifying high risk patients under some conditions (Pourhomayoun and Shakibi, 2021)(Ryan et al., 2020)(Tuti et al., 2017)(Wiemken et al., 2017)(Wu et al., 2014). (Alpaydin, 2016)(Michalski et al., 2013).

Measuring the results of ML algorithms is an essential part of any work in this area. There are several metrics for evaluating performance of a predictive model according to different points of views or needs. Diverse analyses may be accomplished depending on the problem, domain and application at hand. Thus, sometimes considering only one measure to evaluate is not adequate for a given purpose (Hossin and Sulaiman, 2015). For instance, for imbalance class problems, accuracy becomes a poor evaluation measure since it may lead to erroneous conclusions because the model learning tends to classify the majority class.

Receiver Operator Characteristics (ROC) is a bi-dimensional graph commonly used in ML scenarios to analyze and compare classifiers performance. It displays the trade-off between True Positive Rate (TPR)(sensitivity) versus False Positive Rate (FPR)(100 - specificity) at various threshold settings. The higher the ROC passes through the upper left corner, the better the model is able to output correct predictions. On the other hand, the closer the curve comes to the 45-degree diagonal (in the lower right triangle) of the ROC space, the less accurate the test. In medicine, ROC curve plays an important role for clinical decisions towards confirming or not a diagnostic test.

ROC curve provides a way to summarize all of the prediction model information with a focus on the positive class, i.e., the one which is usually object of interest. From a ROC curve it is also possible to extract the Area Under the Curve (AUC), which quantitatively summarizes the ML model performance in the ROC space to a single scalar value, thus enabling to make comparisons among resulting models. AUC takes values from 0 to 1, where value 1 means a perfect classifier which is able to distinguish between all positive and negative class points whereas a value near 0 means a classifier with no ability to discriminate the classes. Decision making in the medical community has an extensive literature on the use of ROC curves for diagnostic testing (Fawcett, 2006). In recent years, there has been an increasing usage of ROC curves by the ML community, due in part to the observation that only simple classification accuracy is often a poor metric for measuring performance of predictive models (Fawcett, 2001)(Fawcett, 2006).

3 RELATED WORKS

Machine Learning techniques have been used in literature to predict mortality risk on patients diagnosed with pneumonia and similar respiratory infections. Some of them are described in the following.

The work of (Wiemken et al., 2017) presents a prediction model of 30-day post discharge mortality on patients diagnosed with pneumonia. The dataset contains a variety of inpatient EMR data, including hourly measurements of vital signs and patient health history. The dataset includes adult patients with no specific age restriction. Experiments show that Naïve Bayes classifier has the best predictive performance for the scenario at hand. Results have been evaluated based on a comparison of performance in terms

of AUC between this work with other related ones. Results indicate an AUC of 0.832 which is better than the compared previous works. This work does not establish a different or actual baseline for comparison. It also suggests as limitation that it is important to evaluate other kinds of features related to pneumonia treatments and also other modeling approaches to improve clinical outcomes.

The XGBoost classifier is evaluated in the work of (Ryan et al., 2020). This work predicts in-hospital mortality up to 72 hours from admission, with focus on data of inpatients on ICU diagnosed with pneumonia, COVID-19 or mechanically ventilated. The prediction models use datasets of patient records collected every 3 hours. The results are compared with mortality risks scores as baselines in classifying patients (qSOFA, MEWS(Long et al., 2017) and CURB-65). Despite presenting AUC values for each risk score at 12-, 24-, 48-, and 72- hour time windows, the work does not provide details regarding the calculation of AUC using the dataset features. Results show the XGBoost classifier surpassing the defined baselines with AUC values of 0.82. 0.81, 0.77 and 0.75 for mortality prediction at 12, 24, 48, and 72 hour time windows. This work focus on a predictive model, which is limited for anticipating patient mortality at specific time points of treatment up to 72 hours. Learning features are restricted to laboratory results and vital signs.

In (Pourhomayoun and Shakibi, 2021), the authors also employ supervised classifier algorithms to predict mortality risk, with focus at triage phase of incoming patients with COVID-19. The dataset includes a total of 112 features of patient EMR data. The main contribution presented by the authors is a feature selection process based on a filter method, highlighting hypertension and age as the most relevant features. The best performance model, obtained by the Random Forest classifier, provided an AUC of 0.94 and a probability for positive classification of up 88%. This work is limited to triage patients and does not present a specific baseline, such as state of the art mortality risk scores. Models results are analysed and compared to each other.

The work of (Tuti et al., 2017) has undertaken a retrospective cohort using clinical characteristics and common comorbidities, w.r.t. increasing risks of inpatient mortality. This work focus on children aged 2–59 months which were admitted with a clinical diagnosis of pneumonia. The evaluated models demonstrate moderate good performance, with the classification algorithm *Partial Least Squares classifier* achieving an AUC of 0.75. Results show that elevated respiratory rates, age ranging from 2 to 11 months and weight-for-age are important features indicating mortality of inpatients. The work findings support the need for re-evaluation of the guidelines for non-severe pneumonia, specifically among infants and in populations where comorbidities are common.

Important points may be discussed from these works. All of them highlight the need to evaluate additional features beyond the ones used in their works in order to try improving predictive models. They focus on mortality prediction w.r.t. different steps of CAP treatment (e.g., 30 days post discharge, triage process or 72 hours of diagnosis). Related works present studies which cover different inpatient groups, such as the ones categorized by age or by the need of hospitalization in ICUs. A common limitation shared in all the works is the absence of a consolidated or actual baseline to analyse prediction model results. While some works propose a baseline evaluation method (Ryan et al., 2020)(Wiemken et al., 2017), no similar description is provided by others. Thus, comparing these works with ours, we may point out some different aspects as follows:

- Our approach focuses on a specific scenario regarding elderly inpatients diagnosed with pneumonia;
- The collected and prepared dataset also includes entry features unique to this study extracted from medical and nursing annotations of patients and their family health history;
- Results include analysis of positive classification probability;
- Comparative analysis performance and statistical test are conducted regarding a learned classifier in comparison with a baseline that is a computational implementation of the CURB-65 score.

4 RESEARCH METHODOLOGY

Diverse data science methodologies have been proposed and used to approach business or research problems (Luo et al., 2021). In this work, we use the CRISP-DM (Cross Industry Standard Process for Data Mining) (Wirth and Hipp, 2000) as a base methodology, since it is one of the most used so far (Schröer et al., 2021). Regarding the health data domain at hand, we have tried to consider some particular issues discussed in the following. Thus, in this section, we describe the applied methodology, which includes steps provided by the CRISP-DM process and also specific steps taken into consideration given this data domain. To this end, we keep in mind the classification problem of this work. Then we present some

particular points and a rationale for the features and baseline used. We also discuss some aspects related to how evaluate results given the context of this work.

4.1 Classification for Death Risk

The risk of mortality from CAP is still a challenge faced by medical teams. This is an even more relevant issue w.r.t. elderly patients, since they are a critical group with increased chances of health complications. Given this context, this work proposes the use of ML based on the CRISP-DM methodology to analyze and predict risk of death w.r.t. elderly inpatients with CAP. We deal with the problem of predicting mortality risk of inpatients with CAP as a binary classification problem. According to some related works research and also to domain specialists, some features that may be related to the death rate due to pneumonia of elderly inpatients have been selected. We define our classification problem as follows.

Suppose that $D_{train} := \{(\vec{x}_i, y_i), \ldots, (\vec{x}_n, y_n)\}$ with $i = 1, \ldots, n$ is a training set, where $\vec{x}_i$ are the feature vectors representing the instances in the feature space $X \in \mathbb{R}^m$, and y_i denotes the class label to which $\vec{x}_i$ belongs to in the set of label for positive $(+)$ and negative $(-)$ classes. The positive class represents the risk of an elder inpatient dying during hospitalization and the negative class indicates the absence of risk. Thus, the purpose of this work is to learn a classification function

$$f : (\vec{x}_i, y_i) \rightarrow \{+, -\}$$

that classifies any given instance on an independent test set D_{test} (not used during training phase) as positive if there is risk of mortality; or negative, otherwise. The prediction function f must minimize the error on D_{test}. It also estimates a probability $\mathcal{P}$ of instances in D_{test} belonging to the predicted class, in a continuous interval $[0 \ldots 1]$.

4.2 Mortality Risk Score

Some assessment tools for evaluating the severity of pneumonia are indeed used in clinical practice. This is helpful since they assist medical decisions on managing outpatient versus inpatient settings in order to optimize hospital referral and lower hospital admission (Pourhomayoun and Shakibi, 2021)(Ryan et al., 2020). Thus, bringing the way in which those assessments are performed in clinical practices to the light of an experimental computational evaluation may produce a more assertive solution.

The reality of the hospital in Brazil that gave rise to this research includes the use of panels with indications based on the CURB-65 score. Therefore, medical professionals take these indications into account in their decisions regarding inpatients with pneumonia. Another score also used is the PSI. However, since data from patients' gasometry exams are not available from the mentioned hospital, it has not been possible to evaluate the PSI score in this work. Therefore, we focus on understanding and applying the CURB-65 score as a baseline to this study.

The CURB-65 score usage is twofold: (i) it has been used to define the features that should be included in the built dataset; and (ii) it has been chosen to build a computational representation of the score as a baseline, according to calculation rules depicted in Section 2. Regarding the former, in addition to data used in the CURB-65 measurement, other features have been acquired or even extracted according to suggestions from domain specialists and also provided by some data understanding. For instance, data w.r.t. patients family health history have also been included. These aspects are presented in the next section.

5 PROPOSED APPROACH

In this section, we present our approach for predicting the risk of death among elder inpatients with CAP. At first, we describe the dataset built and used in our experiments and the tasks which have been accomplished for its preparation. Then, we make some remarks w.r.t. the experimentation scenario, developed baseline and evaluation. In the end, we provide some obtained results and some analyses regarding them.

5.1 Dataset preparation

The dataset collected from a hospital in Brazil is composed by electronic medical records of patients with CAP. Data were gathered from 2018 to 2021, resulting in 64.160 measurement records of 461 elderly patients diagnosed with CAP. Each record contains features that represent measurements of vital signs and laboratory results, taken usually every 3 hours. Personal data of patients were made anonymous during data extraction in order to preserve their privacy.

Based on the CRISP-DM steps (Wirth and Hipp, 2000), data preparation tasks on the originally collected dataset have been performed. To this end, the temporal condition of taking measurements of patients was considered, as well as aspects related to the completeness and correctness of the overall data. The data preparation tasks are described as follows.

a) Data selection: Due to the pandemic of COVID-19 and, since data collected at the hospital

included the period between 2019 and 2021, data initially also contained examples of pneumonia associated with COVID-19. Based on discussions with medical professionals, which are our domain specialists, it has been decided to not consider data specific to patients with COVID-19. This is due to the fact that there are still a lot of misunderstandings and learning around the COVID-19 infection and its association with pneumonia and, particularly, CAP. In order to not disturb the current research, these particular examples have not been considered at data selection.

Features have been mainly defined according to the data needed for calculation of the CURB-65 Score. In addition, domain specialists recommended the inclusion of data provided by patients comorbidities and family health histories. This is rather important since some hereditary diseases could be related to patient conditions and thus might require closer attention during CAP treatment. Other features have also been considered based on limitations provided by some examples discussed in related works, such as the patient family history. Some of them pointed out not so good results due to lack of some attributes, as discussed, for instance, by (Wiemken et al., 2017) in Section 3.

Thus, a set of 30 features has been selected as relevant for the prediction problem at hand, as shown in Table 1. They are categorized as follows: demografic, vital signs, laboratory results, comorbidities, or family health history. Numerical features obtained from patients EMR include: age, hospitalization time (measured in hours), pulse, respiratory frequency, systolic blood pressure, dyastolic blood pressure, temperature, urea nitrogen, sodium, glucose and hematocrit. The categorical features obtained represent the presence or absence of a certain condition w.r.t. a that patient in a given time. Categorical features are as follows: Nursing home resident, smoking history, altered mental state, mechanical ventilation, neoplastic disease, congestive heart failure, cerebrovascular disease, kidney disease, liver disease, cronicle pulmonary disease, cardiovascular disease, psychiatric disease, neurologic disease. In the same way, a feature such as family health history brings cases of diseases which may also be relevant to comprehend a patient diagnosis and evolution (e.g., a neurologic disease). Gender is a categorical feature, but its classification is 0 for male and 1 for female patients as a means of standardization. The understanding of some features and their implications in CAP treatment are not trivial for non-medical people. Nevertheless, as shown in Table 1, features used to build the dataset regard health conditions related to inpatients with CAP. Medical details regarding each one of the conditions are out of this scope.

Table 1: Features of the dataset and their categories

Category	Feature
Demographic Data	Age
	Gender
	Nursing home resident
	Smoking History
Vital Signs	Hospitalization Time(Hours)
	Pulse(bpm)
	Respiratory frequency(bpm)
	Systolic blood pressure(mmHg)
	Dyastolic blood pressure(mmHg)
	Temperature(°C)
	Altered mental state
	Mechanical ventilation
Laboratory Results	Urea nitrogen(mg/dl)
	Sodium(mmol/l)
	Glucose(mg/dl)
	Hematocrit(%)
Comorbidities	Neoplastic Disease
	Congestive Heart Failure
	Cerebrovascular disease
	Kidney disease
	Liver disease
	Cronicle Pulmonary disease
	Cardiovascular disease
	Psychiatric disease
	Neurologic disease
Family History	Neoplastic Disease
	Cardiovascular Disease
	Neurologic disease
	Psychiatric disease

b) Missing values: In order to input values for missing individual patient measurements, we have defined the usage of the average of existing values grouped by a patient identification. Average based construction has been chosen because it disregards outliers and has the least impact on feature values distribution (Hastie et al., 2017). There have been no missing values for the categorical features.

c) Data extraction for categorical features: At hospital, on an initial evaluation of a given patient, the medical and nursing staff fill in an electronic form related to his/her anamnesis. To this end, they select preexisting diseases and/or family health history conditions based on a reference background provided by the hospital (i.e., a semantic dictionary composed by

medical disease terms). Described terms along with some textual descriptions are stored in a specific textual field within the hospital's database. Considering this rich information, some categorical binary features have been extracted in order to provide added features. The rationale used to this end is the following: given a patient set of information provided by a textual field (annotation), if a term, provided by the hospital dictionary, is included in the text, then that term is made a feature of the dataset. This terms follow a pattern of identification used by all the hospital team. Thus, if a patient's annotation contains a given term, which is considered a feature of the specified dataset, its value is set as present (1), otherwise it is defined as absent (0). For instance, if a patient annotation contains the term "Chronicle Pulmonary disease", its namesake feature is set as 1.

Regarding mechanical ventilation and altered mental states, their specific features have been extracted from a boolean field of the hospital database and set on the dataset based on the same method described before: value 1 if the term is present or value 0, otherwise.

d) Labelling: The class labels have been derived according to the following rationale: patient examples which have information regarding the time and cause of their death related to CAP have been labelled with 1 (deceased). Otherwise, patients that remained alive after hospitalization from CAP have been labeled as 0 (survived). After finishing the labelling process and the data preparation tasks described, with respect to the target variable, the dataset is summarized as follows: 43% of positive examples (27,306) and 57% of negative examples (36,854) . This scenario represents the real proportion of data for the period considered at data extraction time.

5.2 Experimental Evaluation

We have conducted some experiments and analyses in the light of our approach. The main experiment has been defined aiming to compare some classification algorithms applied to the intended prediction model at hand. To accomplish this, we have performed a ten times stratified group 10-fold cross-validation (Hastie et al., 2017) using all available data in order to measure the variability of the results. As usual, the models have been induced on the train dataset and have had their performance measured in a test dataset. In this setting, we ensure that instances of the same patient ID are not present in both training and test data, i.e., we avoid ID instance overlapping during training and test.

The classification methods which have been applied in initial experiments were Random Forest (RF), Support Vector Machine (SVM), Multi-Layer Perceptron (MLP) and Logistic Regression (LR). They have been selected based on results and discussions provided by a systematic review of literature (Silva et al., 2020). (Silva et al., 2020) has pointed out that those classifiers were used at state-of-the-art researches regarding pneumonia scenarios due to their applicability and efficiency. All the models have been trained using the default parameters defined in the SciKit-Learn library (Géron, 2019).

As baseline we have chosen to implement a function that computes the risk of death according to the CURB-65 severity score. CURB-65 determines mortality risk estimate based on a subset of four specific features represented by $\vec{s_i} \subset \vec{x_i}$, depicted in Section 2. The baseline evaluation has been developed as follows: for a given feature vector representing the instance $\vec{s_i} \in D_{test}$, and from a given probability score function $curb65(\vec{s_i})$, we assign a positive class to $\vec{s_i}$ if $curb65(\vec{s_i}) \geq thr$, where thr is a pre-specified threshold that represents a score value of 3 points or higher (severe mortality risk). The output provided by $curb65(\vec{s_i})$ also estimates the probability of mortality risk (up to 27.8%).

As mentioned earlier, AUC is the metric used to evaluate the performance of the classification models. In addition, ROC curves have been generated to provide some analysis focused on the most critical classification, i.e., patients correctly classified as at risk of death during treatment. We have computed AUC from prediction scores for both learned classifiers and developed baseline using $roc\_auc\_score()$ method of sklearn library. In addition, we provide the probability of the predictions.

5.3 Results

Table 2 presents the obtained results regarding the comparative evaluation among the classifiers and the baseline. The second column shows the obtained AUC measure. The third to fifth columns point out the minimum, maximum and average probability for each method, which represents the probability of predicting the positive class (patients that will not survive). The probability ranges from 50 to 99% for the evaluated models and from 14% to 27% when considering the CURB-65 baseline. As we can observe in Table 2, all generated classifiers outperform the CURB-65 baseline. The best expected performance in terms of the AUC value (0.81) has been achieved by the Logistic Regression classifier (LR). This indicates a 81% chance that the model correctly distinguishes positive class from negative class, against 61% of the base-

Table 2: Experimentation results considering AUC and probability

Model	AUC	Probability(%)		
		MIN	MAX	AVG
CURB-65	0.61	14	27	20
RF	0.78	50	92	68
SVM	0.71	50	86	75
MLP	0.75	50	98	78
LR	0.81	50	99	78

line. In general, this means that a higher AUC demonstrates the ability of a classifier to identifying more True Positives and Negatives than False Positives and Negatives. Regarding the fifth column of Table 2, the LR classifier is also able to correctly predicting the positive class with 78% of confidence in average, against 20% of CURB-65 score. Therefore, we may point out a promising result for the LR classifier in offering a more reliable estimation for risk of death regarding inpatients with CAP to medical teams.

Figure 1 depicts a ROC curve comparing results obtained by the best classifier (LR) in comparison with the baseline based on the CURB-65 score. The diagonal line from the lower left-hand (0,0) to the upper right-hand (1,1) represents the strategy of a model randomly guessing a class. We are able to observe that there is no intersection between curves. The LR model curve shows better performance, since it is closer to the perfect discrimination (0,1). The curve also demonstrates that the LR model is more conservative than the baseline with CURB-65 since it makes positive classifications only with strong evidences, what implies in reducing the number of false positives. It is worth mentioning that as the threshold gradually increases, LR tends to present better TPR than CURB-65. Even at lowest threshold, the performance of LR is better than the baseline.

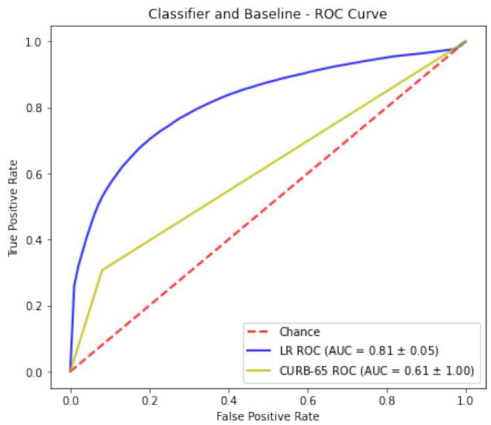


Figure 1: ROC Curve for LR vs CURB-65

Table 3 presents a chronology of patient measurements representing learning features in order to understand the applicability of our approach. In this case study, we consider a patient identified as 354, who is a 95 years old woman, and has no family health history of diseases or comorbidities. This patient did not survive the CAP treatment during her hospitalization. The meaning of acronyms are denoted at the end of the Table 3 (e.g., HHP-Hours hospitalized; RPR-Respiratory Rate). Presented chronological data are based particularly in terms of the hours of hospitalization (HHP) of the patient, which are depicted in the first column of the table. With regards to the first five measurements, it is possible to verify a low probability for risk of death, thus the prediction column has been set with value 0 (a negative classification), although the actual value is 1. After 650 hours of hospitalization, the learned classifier is able to correctly indicate a positive classification. 650 hours of hospitalization means around 30 days. This is explained due to the variation of numeric features of measurements outside their normality values. As an example, taking into account a vital sign pulse (PLS) feature, which usually ranges from 57 to 100 in elderly patients, at last stages of hospitalization, it hits 140, as shown in 3. The usage of mechanical ventilation also indicates an ever increasing probability of death. Further analysis on other features, their correlations and importance to the results of obtained predictions is needed.

Complementary assessments of other inpatients data demonstrate that most of incorrect classifications occurred in cases where time of hospitalization was less than 15 days of treatment. Despite these incorrect classifications, the predictive models in this work show promising results. They also draw attention to patients with more than 700 hours of hospitalization, blood urea higher than 50 or with mechanical ventilation. These patients usually suffer a high decline of survival chances. It is worth mentioning that the CURB-65 baseline has provided even smaller odds of prediction since hospitalization time and most of the numeric features are not part of its calculation.

In order to ensure that the difference of AUC performance between the LR model (best performance) and the current baseline is statistically significant, we have conducted a hypothesis test formulated as follows:

- **H0**: *LR model and CURB65 have the same AUC mean performance* ($\mu_1 = \mu_2$)
- **H1**: *LR model and CURB65 have different AUC mean performance* ($\mu_1 \neq \mu_2$). *Thus, the best performing model outperforms CURB65.*

After performing 10 cross validation runs, each

Table 3: Patient 354 chronology sample and results for the LR model.

HHP	RPR	PLS	SBP	DBP	TMP	HMT	BUR	MCH	Actual	Predict	%
13	21	57	127	64	36.2	28.5	64	0	1	0	23.84%
50	19	70	120	70	37.0	28.2	59	0	1	0	23.01%
164	18	70	130	80	36.4	26.5	50	0	1	0	18.13%
316	18	89	120	80	36.5	26.5	50	0	1	0	30.17%
620	19	67	110	70	36.6	27.6	53	0	1	0	38.29%
650	19	79	120	70	36.5	27.6	53	1	1	0	48.26%
1003	18	87	120	70	36.6	26.9	64	1	1	1	72.93%
2006	19	79	120	80	36.2	25.4	58	1	1	1	87.66%
3044	20	83	130	70	36.5	24.7	40	1	1	1	98.26%
4202	19	75	120	70	36.7	28.2	59	1	1	1	99.76%
4427	18	140	169	79	37.0	24.7	110	1	1	1	99.99%
4955	18	94	118	79	37.0	26.9	64	1	1	1	99.94%
Subtitle:	HHP-Hours hospitalized; RPR-Respiratory Rate; PLS-Pulse; SBP-Systolic Blood Pressure; DBP-Dyastolic Blood Pressure; TMP-Temperature; HMT-Hematocrit; BUR-Blood Urea; MCH-Mechanicly Ventilated;										

one with 10 folds itself, we have had 100 measurements of AUC. The obtained measurements make up the set of samples to be used in a statistic test as presented in Table 4. Each table row refers to a fold in the cross validation process. The means and standard deviation by run are also depicted in Table 4.

Considering that the total sample set includes 100 elements, we have used the Kolmogorov-Smirnov test(Dodge, 2008) in order to verify the set's normality distribution. Accomplished results with the Kolmogorov-Smirnov test demonstrated that data do not differ significantly, thus we can consider them as normally distributed.

Therefore a paired one-tailed *z-test* with 95% confidence has been performed (α= 0.05). By applying the statistical test *Z-test*, a $a = 0.5$ represents a critical value of 1.645 (Davis and Mukamal, 2006) that must be surpassed to which the result enter the distribution zone in which the null hypothesis would be rejected. The computed Z-Value = 40 also represents a probability value of less than 0.00001 at a normal distribution table. Thus, with the Z-Value $>$ 1.645 and *P-Value*(0.0001) $< \alpha$ (0.5), the hypothesis H0 is rejected and we can confirm that the LR model performance is statistically significant better than CURB-65 with α= 0.05.

6 CONCLUSIONS AND FURTHER WORK

Predicting mortality risk with respect to elderly inpatients with CAP is an important issue in hospitals. Based on that issue, we have developed a machine learning approach to classify patients with CAP at risk of mortality during hospitalization. The main purpose is providing means to medical professionals make more assertive decisions. To this end, we also provide higher probability of a positive or negative classification occurs, i.e, our approach is able to indicate inpatients likely to come to death with around 51% to 99%. As a consequence, the presented approach may help increasing elder inpatients being able to survive from CAP. The solution provided by this work includes: (i) an extraction of specific data and features according to the application domain and, particularly, from medical and nurse annotations; (ii) a baseline setting developed according to a real score used in hospitals; (iii) a predictive analysis model which outperforms the defined baseline w.r.t. the AUC metric and (iv) a statistical significance test to further validate the higher performance of the classifier in comparison with the evaluated baseline.

Regarding the classification models evaluation, the obtained results have been compared and analysed. The LR classifier has been able to predict the mortality risk with the best performance by means of the AUC (0.81) metric. It provides an average positive class probability of 78%. The baseline based on the CURB-65 risk score has achieved an AUC of 0.61, with an average probability of 20%. Results have also highlighted that inpatients with more than 30 days at hospital have been classified with significant higher risk of death, what indicates the importance of such feature w.r.t. the classification model. A hypotheses test formulation has confirmed that our approach is statistically significant better than the compared baseline. The results also show some limitations regarding the process of correctly predicting patients at risk of

Table 4: AUC values for repeated 10 times 10-fold cross validation

Run	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Fold										
1	0.73	0.82	0.80	0.80	0.84	0.82	0.83	0.79	0.88	0.80
2	0.72	0.81	0.84	0.73	0.86	0.86	0.78	0.74	0.76	0.84
3	0.83	0.78	0.85	0.81	0.77	0.81	0.80	0.91	0.78	0.84
4	0.89	0.85	0.78	0.89	0.86	0.79	0.85	0.81	0.74	0.76
5	0.78	0.91	0.80	0.80	0.87	0.85	0.84	0.77	0.81	0.73
6	0.79	0.81	0.84	0.71	0.64	0.76	0.73	0.86	0.84	0.78
7	0.74	0.74	0.88	0.89	0.78	0.77	0.85	0.85	0.80	0.76
8	0.84	0.82	0.91	0.76	0.82	0.75	0.84	0.78	0.83	0.82
9	0.80	0.81	0.72	0.85	0.74	0.80	0.77	0.76	0.77	0.79
10	0.80	0.83	0.82	0.79	0.87	0.76	0.87	0.85	0.81	0.73
Mean	**0.79**	**0.81**	**0.82**	**0.80**	**0.81**	**0.80**	**0.82**	**0.81**	**0.80**	**0.79**
Average	**0.81**									
Standard deviation	0.05									

death in early stages of hospitalization. In this situation, the learned classifiers have had not so good performance results.

As future work, we intend to include the gasometry exam results in order to enrich data and enable the experimentation of PSI score as another baseline. In addition, since we have observed a high importance of some features w.r.t. positive classifications and its related probability of death, a detailed feature analysis study will be accomplished. Furthermore, some principles of the methodology and results achieved in this work can be spread out to other kinds of diseases, enabling assistance to health professionals in death risk alerts.

ACKNOWLEDGEMENTS

The authors would like to thank the Alberto Urquiza Wanderley Hospital team and SiDi Technology Institute. Without their support and guidance, it would be impossible to complete this work.

REFERENCES

Alpaydin, E. (2016). *Machine learning: the new AI*. MIT press.

Bezemer, T., de Groot, M., Blasse, E., ten Berg, M., Kappen, T. H., Bredenoord, A. L., van Solinge, W. W., Hoefer, I. E., and Haitjema, S. (2019). A Human(e) Factor in Clinical Decision Support Systems. *Journal of Medical Internet Research*, 21(3):e11732.

Chen, J.-H., Chang, S.-S., Liu, J. J., Chan, R.-C., Wu, J.-Y., Wang, W.-C., Lee, S.-H., and Lee, C.-C. (2010). Comparison of clinical characteristics and performance of pneumonia severity score and curb-65 among younger adults, elderly and very old subjects. *Thorax*, 65(11):971–977.

Davis, R. B. and Mukamal, K. J. (2006). Hypothesis testing: means. *Circulation*, 114(10):1078–1082.

Dodge, Y. (2008). *The concise encyclopedia of statistics*. Springer Science & Business Media.

Fawcett, T. (2001). Using rule sets to maximize roc performance. In *Proceedings 2001 IEEE international conference on data mining*, pages 131–138. IEEE.

Fawcett, T. (2006). An introduction to roc analysis. *Pattern recognition letters*, 27(8):861–874.

Géron, A. (2019). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems*. O'Reilly Media.

Hastie, T., Tibshirani, R., and Friedman, J. (2017). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer.

Hespanhol, V. and Bárbara, C. (2020). Pneumonia mortality, comorbidities matter? *Pulmonology*, 26(3):123–129.

Hossin, M. and Sulaiman, M. (2015). A review on evaluation metrics for data classification evaluations. *International Journal of Data Mining & Knowledge Management Process*, 5(2):1.

Long, B., Long, D., and Koyfman, A. (2017). Emergency medicine evaluation of community-acquired pneumonia: History, examination, imaging and laboratory assessment, and risk scores. *The Journal of Emergency Medicine*, 53(5):642–652.

Luo, E. M., Newman, S., Amat, M., Charpignon, M.-L., Duralde, E. R., Jain, S., Kaufman, A. R., Korolev, I., Lai, Y., Lam, B. D., et al. (2021). Mit covid-19 datathon: data without boundaries. *BMJ innovations*, 7(1).

Michalski, R. S., Carbonell, J. G., and Mitchell, T. M. (2013). *Machine learning: An artificial intelligence approach*. Springer Science & Business Media.

Pourhomayoun, M. and Shakibi, M. (2021). Predicting mortality risk in patients with covid-19 using machine learning to help medical decision-making. *Smart Health*.

Ryan, L., Lam, C., Mataraso, S., Green-Saxena, A. A. A., and Pellegrini, E. (2020). Mortality prediction model for the triage of covid-19, pneumonia, and mechanically ventilated icu patients: A retrospective study. *Annals of Medicine and Surgery*.

Schröer, C., Kruse, F., and Gómez, J. M. (2021). A systematic literature review on applying crisp-dm process model. *Procedia Computer Science*, 181:526–534.

Silva, V., Novo, A. D. R., Souza, D., and Rêgo, A. (2020). Machine learning to assist in pneumonia decision making: A systematic review of the literature. In *Anais do VIII Symposium on Knowledge Discovery, Mining and Learning*, pages 201–208. SBC.

Tuti, T., Agweyu, A., Mwaniki, P., Peek, N., and English, M. (2017). An exploration of mortality risk factors in non-severe pneumonia in children using clinical data from kenya. *BMC medicine*, 15(1):1–12.

Webb, A. and Gattinoni, L. (2016). *Oxford Textbook of Critical Care*. Oxford University Press.

Wiemken, T., Furmanek, S., Mattingly, W., Guinn, B., and Cavallazzi, R. (2017). Predicting 30-day mortality in hospitalized patients with community-acquired pneumonia using statistical and machine learning approaches. *Journal of Respiratory Infections*.

Wiemken, T., Kelley, R., and Ramirez, J. (2013). Clinical scoring tools: which is best to predict clinical response and long-term outcomes? *Infectious disease clinics of North America*.

Wirth, R. and Hipp, J. (2000). Crisp-dm: Towards a standard process model for data mining. In *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining*, volume 1. Springer-Verlag London, UK.

World Health Organization (2015). Health Topics: Pneumonia. https://www.who.int/health-topics/pneumonia.

Wu, C., Rosenfeld, R., and Clermont, G. (2014). Using data-driven rules to predict mortality in severe community acquired pneumonia. *PLoS One*, 9(4):e89053.

Wu, D., Wu, C., Zhang, S., and Zhong, Y. (2019). Risk factors of ventilator-associated pneumonia in critically iii patients. *Frontiers in pharmacology*, 10:482.