UNIVERSIDADE FEDERAL DE SANTA CATARINA
CAMPUS ARARANGUÁ
PROGRAMA DE PÓS-GRADUAÇÃO EM TECNOLOGIA DA INFORMAÇÃO E COMUNICAÇÃO

Jefferson Pacheco dos Santos

PROPOSTA DE UM SISTEMA PARA AVALIAÇÃO DE RISCOS DE INFECÇÃO DO SÍTIO CIRÚRGICO UTILIZANDO TÉCNICAS DE INTELIGÊNCIA ARTIFICIAL

Araranguá
2021

Jefferson Pacheco dos Santos

PROPOSTA DE UM SISTEMA PARA AVALIAÇÃO DE RISCOS DE INFECÇÃO DO SÍTIO CIRÚRGICO UTILIZANDO TÉCNICAS DE INTELIGÊNCIA ARTIFICIAL

Dissertação submetida ao Programa de Pós-graduação em Tecnologia da Informação e Comunicação da Universidade Federal de Santa Catarina para a obtenção do título de Mestre em Tecnologia da Informação e Comunicação
Orientadora: Prof. Dra. Eliane Pozzebon.
Coorientador Prof. Dr. Antonio Carlos Sobieranski

Araranguá
2021

Ficha de identificação da obra elaborada pelo autor,
através do Programa de Geração Automática da Biblioteca Universitária da UFSC.

Santos, Jefferson Pacheco do
PROPOSTA DE UM SISTEMA PARA AVALIAÇÃO DE RISCOS DE INFECÇÃO DO SÍTIO CIRÚRGICO UTILIZANDO TÉCNICAS DE INTELIGÊNCIA ARTIFICIAL / Jefferson Pacheco do Santos ; orientador, Eliane Pozzebon, coorientador, Antonio Carlos Sobieranski, 2021.
139 p.

Dissertação (mestrado) - Universidade Federal de Santa Catarina, Campus Araranguá, Programa de Pós-Graduação em Tecnologias da Informação e Comunicação, Araranguá, 2021.

Inclui referências.

1. Tecnologias da Informação e Comunicação. 2. Inteligência Artificial. 3. Mineração de Dados. 4. Algoritmo. 5. Infecção do Sitio Cirúrgico. I. Pozzebon, Eliane . II. Sobieranski, Antonio Carlos . III. Universidade Federal de Santa Catarina. Programa de Pós Graduação em Tecnologias da Informação e Comunicação. IV. Título.

Jefferson Pacheco dos Santos

PROPOSTA DE UM SISTEMA PARA AVALIAÇÃO DE RISCOS DE INFECÇÃO DO SÍTIO CIRÚRGICO UTILIZANDO TÉCNICAS DE INTELIGÊNCIA ARTIFICIAL

O presente trabalho em nível de mestrado foi avaliado e aprovado pela banca examinadora composta pelos seguintes membros:

Prof. Roderval Marcelino, Dr.
Instituição Universidade Federal de Santa Catarina – UFSC

Prof. Antônio dos Reis Sá Dr.
Universidade Federal de Santa Catarina – UFSC

Profª Josete Mazon Dr.(a)
Universidade Federal de Santa Catarina - UFSC

Certificamos que esta é a **versão original e final** do trabalho de conclusão que foi julgado adequado para obtenção do título de mestre em Tecnologia da Informação e Comunicação.

Documento assinado digitalmente
Juarez Bento da Silva
Data: 06/10/2021 08:56:32-0300
CPF: 366.921.580-87
Verifique as assinaturas em https://v.ufsc.br

Coordenação do Programa de Pós-Graduação

Documento assinado digitalmente
Eliane Pozzebon
Data: 06/10/2021 06:32:28-0300
CPF: 779.213.139-49
Verifique as assinaturas em https://v.ufsc.br

Profª Eliane Pozzebon, Dra.
Orientadora

Documento assinado digitalmente
Antonio Carlos Sobieranski
Data: 06/10/2021 09:58:01-0300
CPF: 005.305.809-77
Verifique as assinaturas em https://v.ufsc.br

Prof. Antonio Carlos Sobieranski, Dr.
Coorientador

Este trabalho é dedicado a toda a minha amada família.

AGRADECIMENTOS

Agradeço a Deus por me dar coragem, paciência, perseverança e proteção em todos os momentos durante o mestrado, onde em meio a tantas funções e responsabilidade que me dividiam numa rotina intensa de trabalho, família, estudos e pesquisas. Só tenho a agradecer e muito!

A minha querida esposa e parceira, Kelimar Bergman, por sempre me auxiliar nos momentos mais difíceis desta jornada.

À minha família, pais, irmã, sogro e sogra que além de auxiliar nos cuidados com minha filha Julia Bergman dos Santos para que eu pudesse me dedicar aos estudos, me incentivaram em todo o processo do mestrado, desde os estudos das disciplinas isoladas até conclusão do mesmo, com palavras de motivação, incentivo e muito carinho.

A três colegas de mestrado, Arildo Sônego que me incentivou a iniciar o mestrado, Reginaldo José da Rosa e Rodolfo Faquin Della Justina, pessoas que marcaram minha trajetória com muitas risadas, desabafos e trocas de conhecimento. Pessoas cruciais para que o objetivo fosse alcançado. Obrigado por me ajudarem sempre nas minhas várias dúvidas e dificuldades.

À minha orientadora Eliane Pozzebon e coorientador Antonio Carlos Sobieranski, pela confiança depositada em mim e nos meus estudos. Vocês são profissionais incríveis, sempre nos orientando e motivando durante a jornada acadêmica. Obrigado professores!

Ao colega Bruno Bitencourt Luiz, Cientista de Dados que tem um coração que não cabe dentro de si, que não mediu esforços durante o desenvolvimento da aplicação de minha dissertação, me auxiliando em cada etapa.

À Universidade Federal de Santa Catarina, aos colegas de mestrado, aos colegas do LabTec, e aos professores do PPGTIC por cada momento especial de aprendizado durante estes anos de dedicação!

“A verdadeira motivação vem de realização, desenvolvimento pessoal, satisfação no trabalho e reconhecimento” (Frederick Herzberg).

RESUMO

Na realidade hospitalar ter um sistema que possibilite a análise de risco de Infecção de Sítio Cirúrgico (ISC) de um paciente antes dele se submeter a um procedimento de saúde, traz segurança ao paciente, pois reforça as medidas preventivas. Para o hospital tem por objetivo a preservação da vida e redução dos custos, reduzindo o impacto financeiro, por consequência, trazendo novas perspectivas e melhorias a saúde. O objetivo deste trabalho consiste no desenvolvimento de uma ferramenta que irá coletar e analisar informações, com o apoio de técnicas de mineração de dados e inteligência artificial, relacionadas às Infecções do Sítio Cirúrgico (ISC) em uma base de dados de pacientes do hospital UNIMED de Criciúma, Sul de Santa Catarina. Esta análise se dará de forma automatizada e se baseará em dados existentes no prontuário/histórico do paciente e que foram analisadas e tratadas de forma estatística, gerando um modelo representação de comportamento, que pode ser integrado a um sistema existente na unidade. Por meio desta ferramenta, é possível que um treinamento contínuo permita um aprimoramento contínuo com base na atualização de comportamentos. Para obtenção dos resultados foram utilizados 07 (sete) algoritmos para testes e o melhor resultado obtido se deu por 02 (dois) algoritmos onde apresentou um acerto de quase 80% na detecção dos perfis de infecção, convertendo em prevenção, evitando riscos maiores para o paciente e, certamente, gerar economia devido aos custos que diante de um quadro de infecção se tornam elevados. Como resultados têm-se dados que demonstram que a Inteligência Artificial utilizada no desenvolvimento de uma ferramenta pode contribuir de forma positiva na área da saúde. Sendo assim, pode-se perceber a contribuição da ferramenta, onde além da prevenção de infecções no sítio cirúrgico, os gastos médios podem ser reduzidos em cerca da 01 (um) milhão de reais com infecções pós-cirúrgicas anualmente apenas no hospital onde foram realizados os estudos.

Palavras-chave: Inteligência Artificial. Mineração de Dados. Algoritmo, Infecção do Sitio Cirúrgico.

ABSTRACT

In the hospital reality, having a system that makes it possible to analyze the risk of Surgical Site Infection (SSI) of a patient before undergoing a health procedure, brings safety to the patient, as it reinforces preventive measures. For the hospital, it aims to preserve life and reduce costs, reducing the financial impact, therefore, bringing new perspectives and improvements to health. The objective of this work is to develop a tool that will collect and analyze data, with the support of data mining and artificial intelligence techniques, related to Surgical Site Infections (SSI) in a patient database of a hospital UNIMED of Criciúma, of the South of Santa Catarina. This analysis will be done in an automated way and will be based on existing data in the patient's medical record/history and that were analyzed and treated in a statistical way, generating a behavior representation model, which can be integrated to an existing system in the hospital. Tich this tool, it is possible that continuous training allows continuous improvement based on updating behaviors. To obtain the results, 07 (seven) algorithms were used for tests and the best result was obtained by 02 (two) algorithms, which showed a success of almost 80% in the detection of infection profiles, converting into prevention, avoiding greater risks for the patient and, certainly, generate savings due to the costs that, when faced with an infection, becomes high. Hence, there are evidences that the artificial Intelligence used in the development of a tool can contribute positively in the health area. thus, one can see the por it is possible to confirm the contribution of the tool, where in addition to the prevention of infections in the surgical site, the average expenses can be reduced by about 01 (one) million BRL with post-surgical infections annually only in the hospital where the studies were performed.

Keywords: Artificial Intelligence. Data Mining. Algorithm. Surgical Place Infection.

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ASP - Programas de Administração Antimicrobiana

AVC - Acidente Vascular Cerebral

CCIH - Comissão de Controle de Infecção Hospitalar

CDC - Centers for Disease Control and Prevetion

CDSS - Sistema de Apoio à Decisão Clínica

DARPA - Defense Advanced Research Projects Agency

DART - Dynamic Analysis and Replanning

DEIS - Departamento de Eletrônica, Informática e Sistemática

DMSS - Sistema de Vigilância de Mineração de dados

EDA - Exploratory Data Analysis

EHRs - Electronic Health Records

EUA - Estados Unidos da América

GPS - General Problem Solver

HACs - Complicações Adquiridas em Hospitais

HAIs - Healthcare-Associated Infections

HISs - Hospital Information Systems

IA - Inteligência Artificial

ICPs - Profissionais de Controle de Infecção

IG - Ganho de Informação

ISC - Infecção de Sítio Cirúrgico

IRAS - Infecção Relacionada à Assistência à Saúde

ITU - Infecções do Trato Urinário

LAD - Least Absolute Deviation

LISs - Sistemas de Informações de Laboratório

LT - Logic Theorist

MISs - Medical Information Systems

MIT - Massachusetts Institute of Technology

ML - Machine Learning

NASA - National Aeronautics and Space Administration

NCCLS - National Committee for Clinical Laboratory Standards

NIM - Marcador de Infecção Hospitalar Eletrônico

OCR - Optical Character Recognition

PDMSs - Sistemas de Gerenciamento de Dados de Pacientes

TAL - Technology Availability Level

TIC - Tecnologias de informação e comunicação

TC - Tomografia Computadorizada

SUMÁRIO

1 INTRODUÇÃO

A cada dia um número crescente de pessoas, em variados segmentos, torna-se dependente do uso contínuo de novas tecnologias, principalmente das Tecnologias de informação e comunicação (TIC). O avanço das TICs permitiu novas formas de comunicação, que podem facilitar a interação entre os indivíduos de uma sociedade.

É comum a vida em sociedade com as facilidades que as tecnologias da informação e comunicação têm a oferecer. Seja na vida pessoal ou profissional, a computação se tornou uma comodidade, agindo, inclusive, como uma extensão dos sentidos (GRANVILLE, 2019).

Tal relevância estende-se ao ambiente hospitalar, onde as TICs apresentam um conjunto de técnicas e ferramentas destinadas à produção de informação gerencial e a descoberta de conhecimentos em grandes bases de dados. Estas técnicas, aplicadas aos dados podem representar um avanço substancial e contribuir, decisivamente, nos estudos referente a Infecção do Sítio Cirúrgico (ISC) através da identificação e correlação de padrões existentes nos dados.

A Infecção de Sítio Cirúrgico é definida como a infecção ocorrida no local do procedimento cirúrgico e está relacionada consequentemente a partir de uma complicação local da região cirúrgica. É a resposta inflamatória provocada pela invasão ou presença de micro-organismos em tecidos orgânicos (SOUZA; MOZACHI, 2009). Na concepção de Carvalho *et al.* (2017) a Infecção Relacionada à Assistência à Saúde (IRAS) é objeto de grande preocupação dos serviços de saúde. Dentre as topografias das IRAS, a Infecção de Sítio Cirúrgico está diretamente relacionada aos procedimentos cirúrgicos, sendo, atualmente, uma das mais importantes entre as IRAS.

Para minimizar este inconveniente, sugere-se como alternativa a descoberta de conhecimento associado à infecção do sítio cirúrgico em bases de dados de pacientes através do uso de técnicas de Mineração de Dados e Inteligência Artificial, com isso contribuindo para identificação e controle de infecção do sítio cirúrgico.

Como descrevem Ramesh *et al*. (2004) o impacto da Inteligência Artificial na medicina pode ser observado em três níveis: (1) para os profissionais de saúde, especialmente por intensificar uma análise de imagens de forma mais rápida e precisa; (2) para o sistema de saúde, reduzindo os erros médicos e modernizando o fluxo de trabalho; e (3) para o paciente, uma vez que, possibilita o processamento das informações deste paciente visando a promoção de saúde.

Além disso, a Inteligência Artificial na medicina é utilizada por grandes empresas como a *IBM*, como uso de determinadas plataformas especializadas, como, por exemplo, o *Watson Health* (SANTOS; VECHIO, 2020). Um exemplo da aplicação do *Watson Health,* como relata Lobo (2018), é a análise realizada com cerca de 35 mil imagens de retina, o resultado demonstra uma acurácia de 80% no diagnóstico de retinopatia, com base na tecnologia *EyePACs*, a qual identifica lesões e outras manifestações percebidos em vasos sanguíneos. Nos EUA um número expressivo de instituições de saúde utilizam o sistema *Watson*, e aproximadamente 80% das informações médicas estão armazenadas no banco de dados do *Watson.*

Bem como, a inteligência Artificial vem sendo aplicada na saúde por métodos 3D com o objetivo de auxiliar na área de radiologia, histologia e oftalmologia identificando e medindo a estrutura de tecidos patológicos. Assim como, a inteligência artificial pode ser utilizada na identificação e monitoramento de casos de insuficiência cardíaca. Enquanto na área terapêutica, a robótica se destaca tanto na execução de cirurgias, quanto na urologia, além de gerar ferramentas inteligentes para terapia intensiva. No futuro, a inteligência artificial será utilizada em todas as áreas da medicina que admitam uma padronização apropriada (CALEGARI, 2019).

No Brasil, como menciona Ribeiro (2020), o hospital Israelita *Albert Einstein* em São Paulo possui aparelhos de imagens que atuam sem a interferência humana. Estes aparelhos podem identificar anomalias em testes e os médicos são notificados automaticamente pelo sistema para que possam tomar decisões rapidamente e com mais acurácia.

Ribeiro (2020) ainda acrescenta que em 2020, em decorrência da pandemia do coronavírus (Covid – 19), os especialistas do hospital Israelita *Albert Einstein* em conjunto com os cientistas do laboratório Labdaps da Universidade de São Paulo (USP-SP), desenvolveram um algoritmo exclusivo de IA capaz de cruzar informações, como resultados de exame de sangue com dados de internação de pacientes na unidade hospitalar, com objetivo de auxiliar a detecção de pacientes contaminados pela COVID – 19. O método foi treinado, entre os dias 17 e 30 de março de 2020, com os dados de aproximadamente 235 pacientes suspeitos de contaminação pelo coronavírus, sendo que destes, 102 pacientes testaram positivos para o coronavírus.

Nesta mesma linha de considerações, Margarido (2020) destaca que, no que diz respeito a IA utilizada no sistema de saúde, o maior beneficiamento consiste nas ferramentas

preditivas que podem presumir o risco de readmissão do paciente no hospital, como o risco de sépsis ou quais pacientes se beneficiariam de cuidados paliativos. Isto pode ser baseado no tratamento da informação disponível nos registros médicos digitais. Assim como, o processamento da linguagem natural pode ser fundamental nesta área, bem como, pode ser útil na tentativa de substituir os teclados e a escrita humana em visitas médicas.

Rouhiainen (2018) aponta que, a IA é a habilidade da máquina de usar algoritmos para aprender e usar o que aprendeu para tomar decisões como um humano, porém com a possibilidade de analisar um massivo número de informações instantaneamente.

Enquanto o Stanford Report de 2016 aponta que o uso de inteligências artificiais tem causado impacto em diversas áreas como: transporte, saúde, segurança, indústria, entretenimento, entre outros. Na educação a estimativa é de que em países como os Estados Unidos, nos próximos quinze anos, o uso de tutores inteligentes e outras tecnologias de IA tanto para auxiliar professores como para o ensino aos alunos se expanda de forma significativa (STONE *et al*., 2016).

No tocante a mineração de dados, de acordo com Gandhi e Tandon (2019), pode ser considerada uma área interdisciplinar que abrange técnicas de máquinas de conhecimentos, reconhecimentos de padrões, estatísticas, banco de dados e visualização, com objetivo de extrair informações de grandes bases de dados, e na área da saúde este processo é fundamental para efetuar diagnósticos e o reconhecimento de padrões de eventos regulares.

Para Fayyad, Piatetsky-Shapiro e Smyth (1996) o objetivo do processo de mineração é fornecer informações que possibilitem montar melhores estratégias de marketing, vendas, suporte, melhorando assim os seus negócios, e isso é feito por etapas que possuem atividades em que são necessárias algumas escolhas de técnicas importantes para o processo, sendo que as decisões tomadas nessas fases e as escolhas das técnicas e algoritmos podem interferir no resultado final do processo.

Bem como, Hand, Mannila e Smyth (2001) relatam que, o objetivo da mineração de dados consiste em encontrar princípios entre as variáveis e padrões, bem como a extração de conhecimento de dados analisados, isto significa que, a mineração de dados busca a descoberta de informações. Entre as técnicas utilizadas para mineração de dados, descritas na literatura, têm-se: Árvores de decisão e Redes Neurais.

Como definem Carvalho, Escobar e Tsunoda (2014) a mineração de dados representa uma opção para o processamento de grandes volumes de dados dos sistemas de dados na

saúde, uma vez que é capaz de encontrar padrões adequados, novos e impressionantes, oferecendo suporte em análises complexas referentes aos dados clínicos.

Neste sentido, Banaee, Ahmed e Loutfi (2013) identificam três tipos de técnicas principais na mineração de dados aplicadas, especialmente, referentes aos dados de saúde: (1) A predição vastamente utilizada com o propósito de identificar fatos que ainda não aconteceram; (2) detecção de anomalias que auxilia na identificação de padrões anormais, os quais não correspondem ao comportamento previsto em determinados grupos de dados e (3) tomada de decisões para diagnósticos derivada da detecção incomuns em dados da saúde que proporciona ao profissional de saúde decisões mais precisas .

1.1 PROBLEMÁTICA E JUSTIFICATIVA

As Tecnologias de Informação e Comunicação estão em constante evolução, presentes cada vez com maior frequência na vida moderna e tornando-se elementos onipresentes aos seres humanos. Diante do crescimento exponencial das informações, as técnicas de mineração de dados e Inteligência Artificial auxiliam na extração de conhecimento, servindo de apoio à tomada de decisões em seus mais diversos patamares.

No cenário hospitalar atual fica claro perceber o índice considerável de risco de infecção pós-cirúrgica, onde traz consigo uma série de preocupações e problemas tanto para o hospital quanto para o paciente, que de um lado geram transtornos como aumento de estadia hospitalar, que aumentam as chances de novas infecções, risco de novos procedimentos cirúrgicos e risco de óbito, do outro lado um alto custo para tratar e recuperar a vida do paciente com a utilização de medicações, leito, profissionais qualificados, gerando assim um custo que se pudesse ser evitado, teria como investir em melhorias com tecnologia e equipamentos hospitalares. A Inteligência Artificial juntamente com a Mineração de dados são tecnologias que vem propor meios de identificar preventivamente possíveis infecções diante de um procedimento cirúrgico, permitindo assim uma postura médica que através de protocolos conseguirá se posicionar de maneira a evitar infecções posteriores.

Com relação a mineração de dados ou Data Mining, é uma etapa fundamental no processo de Descoberta de Conhecimento em Bases de Dados, onde recebe grande destaque na literatura. Como afirmam Fayyad, Piatetsky-Shapiro e Smyth (1996), mineração de dados é o processo de reconhecimento de padrões válidos ou não, existentes nos dados armazenados em grandes bancos de dados. Já conforme Linoff e Berry (2011) mineração de dados é a

exploração e análise, de forma automática ou semi-automática, de grandes bases de dados com objetivo de descobrir padrões e regras.

A mineração de dados é realizada por meio de diversas tarefas, entre elas, a classificação, a associação e o agrupamento. A classificação tem como objetivo classificar itens de acordo com análises previamente realizadas. A associação consiste em descobrir todas as associações em que a presença de um conjunto de itens em uma transação implica em outros itens. O agrupamento separa os dados em vários grupos, de acordo com a similaridade destes dados (PANSONATO; TOMAZELA, 2014).

A ISC leva a graves consequências, incluindo o aumento nos gastos devido ao seu tratamento e a um aumento do tempo de internação. O risco de morte dos pacientes com ISC mostra-se aumentado quando comparado aos que não desenvolveram a infecção.

As graves consequências impostas aos pacientes que desenvolveram a ISC determinam a necessidade de dedicar esforços para a criação de estratégias para a prevenção dessa infecção. Uma das estratégias utilizadas é a determinação de fatores de risco, o que permite identificar situações ou condições clínicas que predisponham ao desenvolvimento da ISC. Neste sentido, a identificação dos fatores de risco para a ISC contribui para a adoção precoce de intervenções de enfermagem que objetivam minimizar esse tipo de complicação pós-operatória.

Estas afirmativas vêm de encontro ao exposto anteriormente, a respeito da atenção dedicada nestes procedimentos. Logo, em um ambiente hospitalar, a possibilidade de automatizar este processo, representaria um avanço tecnológico significativo. Desta maneira, sugere-se a indagação principal desta proposta: **Quais resultados a inteligência artificial traria para a prevenção desta classe de síndrome infecciosa?**

Neste contexto, a proposta desta pesquisa é buscar conhecimento em bases de dados de pacientes do hospital UNIMED de Criciúma do Sul de Santa Catarina, com o apoio de técnicas de inteligência artificial, objetivando compreender possíveis causas de infecções hospitalares, traçando um panorama que possa servir de apoio à prevenção desta classe de síndrome infecciosa.

1.2 OBJETIVOS

1.2.1 Objetivo Geral

O intuito principal deste trabalho diz respeito ao desenvolvimento de uma ferramenta que irá coletar e analisar informações, com o apoio de técnicas de mineração de dados e inteligência artificial, relacionadas às Infecções do Sítio Cirúrgico (ISC) em uma base de dados de pacientes do hospital UNIMED de Criciúma no Sul de Santa Catarina.

1.2.2 Objetivos Específicos

Para atingir o objetivo geral, foram necessárias algumas etapas que são elencadas a seguir:

- Elaborar uma revisão da literatura com conceitos, referências e exemplos para justificar e apresentar conhecimentos sobre o objeto de estudo;
- Realizar um estudo sobre infecções do sítio cirúrgico, suas causas e consequências;
- Identificar, estudar e aplicar técnicas de IA e algoritmos de mineração de dados que possam auxiliar na tarefa de extração de conhecimento;
- Coletar, analisar e apresentar os resultados obtidos;
- Desenvolver uma ferramenta que possibilite contribuir para identificação e controle de infecção do sítio cirúrgico.

1.3 METODOLOGIA ADOTADA

Como descrevem Prodanov e Freitas (2013), existem diversas formas de classificar uma pesquisa. No entanto, para classificação dos tipos de pesquisas é necessário enfatizar que são adotados critérios e estes podem mudar de acordo com o foco da abordagem, os objetivos, os campos, as metodologias, as situações e os objetos de estudo. Neste sentido Kauark, Manhães e Medeiros (2010, p. 25) afirmam que “A importância de conhecer os tipos de pesquisas existentes está na necessidade de definição dos instrumentos e procedimentos que um pesquisador precisa utilizar no planejamento da sua investigação".

Da mesma forma, Gil (2010) acrescenta que a condução de uma pesquisa é realizada por meio dos conhecimentos disponíveis, bem como, a aplicação minuciosa de métodos, técnicas e outros procedimentos científicos, no transcorrer um processo que compreender várias etapas, que tem inicio na definição correta do problema e termina com a apresentação aceitável dos resultados.

Sendo assim, a classificação exposta nesta pesquisa está fundamentada na categorização de acordo com a finalidade, objetivos e abordagem, proposta por Gil (2010). Como também, baseia-se em procedimentos metodológicos descritos por Marconi e Lakatos (2007).

No que diz respeito à sua abordagem, esta dissertação tem como foco uma visão qualitativa dos resultados obtidos na empresa objeto de estudo para compreender como a aplicação da IA pode contribuir para prevenção de infecções no sítio cirúrgico. Como abordam Silva e Menezes (2005), a pesquisa qualitativa leva em conta que existe uma relação dinâmica entre o mundo real e o sujeito, em que não é necessário o uso de métodos e técnicas estatísticas, uma vez que o ambiente natural é a origem direta para coleta de dados e o pesquisador é o instrumento-chave, visto que o processo e seu significado são os motivos principais.

Quanto aos procedimentos técnicos, a pesquisa é classificada como experimental. Conforme Gil (2010), após a definição do objeto de estudo, definem-se as variáveis que podem influenciá-lo, determinam-se os métodos de controle e de observação dos resultados que a variável gera no objeto.

Quanto aos objetivos, a pesquisa pode ser classificada como exploratória uma vez que busca agregar conhecimentos ao pesquisador sobre o tema, baseando-se em fontes primárias como revisão bibliográfica em livros, artigos, dissertações e teses voltadas para compreensão da Inteligência Artificial aplicada à saúde, *Machine Learning* e mineração de dados. Segundo Gil (2010, p. 44), "A pesquisa bibliográfica é desenvolvida com base em material já elaborado [...]", que podem servir de base teórica para esta pesquisa.

No que diz respeito à coleta de dados, esta pesquisa contará com elementos numéricos e textuais derivados de registros extraídos com a mineração de dados para composição de um conjunto de dados qualitativos para realização de análise, com propósito de buscar maior confiabilidade nos resultados apresentados. Como enfatiza Creswell (2007, p. 18), "A utilização de amostras intencionais, a coleta de dados com perguntas abertas, as análises de texto ou imagens, a representação da informação em gráficos e tabelas, e a interpretação pessoal dos resultados das averiguações, todas constituem subsídios aos procedimentos qualitativos".

Como procedimentos metodológicos sugerem-se as seguintes etapas:

1. Delimitação do escopo do trabalho;
2. Levantamento bibliográfico;

3. Estudo e compreensão das técnicas de mineração de dados;
4. Escolha de ferramentas para a definição do modelo proposto;
5. Desenvolvimento da aplicação proposta;

1.4 ESTRUTURA DA DISSERTAÇÃO

Esta pesquisa está dividida em cinco capítulos, estruturados da seguinte forma:

O primeiro capítulo apresenta a introdução, a problemática e justificativa, os objetivos, tanto geral, quanto específicos, a metodologia adotada, a estrutura da dissertação, a delimitação da pesquisa e aderência ao PPGTIC e linha de pesquisa do programa.

O segundo capítulo expõe a inteligência artificial, refere-se a Machine Learning, a inteligência artificial aplicada a saúde. Assim como, retrata a infecção relacionada à assistência à saúde, a infecção do sítio cirúrgico, a análise exploratória de dados, da limpeza de dados e Feature Engineering, e ainda sobre estatística descritiva e estatística inferencial, algoritmos de árvore, classificação e regressão, algoritmos de classificação, treinamento e otimização de algoritmos, métricas de performance de algoritmos e por fim, trabalhos correlatos.

No terceiro capítulo é exposto o projeto ISC (Infecção do Sítio Cirúrgico), apresentação da proposta, os procedimentos de desenvolvimento, os resultados obtidos e análise do problema.

No quarto capítulo é apresentada a análise e discussão dos resultados obtidos na pesquisa.

Por fim, no quinto capítulo é exibida a conclusão da pesquisa, demonstrando as considerações finais do pesquisador e sugestões para trabalhos futuros visando melhorar e aprofundar.

Como elementos pós-textuais são apresentadas as referências utilizadas no desenvolvimento do trabalho, seguidos dos apêndices e anexos. Tendo como Apêndice A, a carta de aceito, Apêndice B a tabela de procedimentos, e como Anexos A, a proposta de classificação da inteligência artificial e aplicações mediadas por IA na Medicina e na Atenção à Saúde de acordo com seu caráter benéfico versos prejudiciais. SW = software, AR = realidade aumentada, VR = realidade virtual, IoT = INTER.

1.5 DELIMITAÇÃO DA PESQUISA

Esta pesquisa aborda a avaliação de riscos de infecção do sítio cirúrgico com a utilização de técnicas de inteligência artificial. Como enfatizam Braga *et al*.(2018, p. 937), "A inteligência artificial (IA) é um campo da ciência da computação que imita os processos de pensamento humano, a capacidade de aprendizagem e o armazenamento de conhecimento". Neste sentido Lobo (2018), a IA é capaz de seguir algoritmos de decisão estabelecidos por especialistas, ter a capacidade de compreender conceitos e não somente realizar o processamento de dados, mas obter raciocínios derivados da capacidade de agregar novas experiência e por consequência se auto aprimorar (*self-learning*) para resolução de problemas e realização de tarefas.

Como também, discutirá sobre o uso de mineração de dados para coleta de informações referentes aos dados sobre infecção do sítio cirúrgico. Segundo Carvalho (2018), a mineração de dados pode fornecer acesso a conhecimento oculto em grandes quantidades de informações que podem gerar melhores recursos tanto para os profissionais de saúde tratar os pacientes, como também, para que os gestores consigam planejar de forma mais adequada os investimentos, ou até mesmo, a prevenção de endemias que venham a acontecer.

Além disso, a pesquisa tem enfoque na escolha de uma ferramenta para que seja possível o desenvolvimento da aplicação proposta para que um ambiente seja disponibilizado para os profissionais da área da saúde, visando a compreensão sobre infecções do sítio cirúrgico e como evitar estas infecções.

A natureza exploratória, experimental e qualitativa da pesquisa é justificável em razão da limitação do tempo do trabalho, devido à complexidade do tema, bem como, para melhor compreensão do tema, de suas variáveis e demais desafios que podem surgir no decorrer da pesquisa. Desta forma, pode-se perceber a importância deste estudo, pois o mesmo pode contribuir de forma efetiva para o objetivo deste trabalho.

1.6 ADERÊNCIA AO PPGTIC E A LINHA DE PESQUISA

A presente dissertação tem foco utilização de Inteligência Artificial e mineração de dados visando contribuir para o desenvolvimento de uma ferramenta para avaliar os riscos de infecção na área da saúde.

Portanto, por se tratar de um processo para o desenvolvimento desta tecnologia da informação e comunicação, a pesquisa enquadra-se no foco do Programa de Pós-Graduação em Tecnologia da Informação e Comunicação (PPGTIC), e na linha de pesquisa Tecnologia

Computacional, a qual visa “[...] desenvolver modelos, técnicas e ferramentas computacionais auxiliando na resolução de problemas de natureza interdisciplinar” (UNIVERSIDADE FEDERAL DE SANTA CATARINA, 2021).

Neste sentido, conforme enfatiza Tamayo (2003), a interdisciplinaridade se refere a uma metodologia de pesquisa científica que tem por objetivo favorecer a integração de resultados de diversas disciplinas, sendo esta decorrente duas origens: (1) Interna, a qual está direcionada em uma mudança do sistema científico, juntamente com seu progresso e organização, e (2) Externa que é representada pela crescente mobilização em prol do saber e proliferação de especialistas.

Sendo assim, esta pesquisa delimita-se nesta linha de pesquisa do PPGTIC, pois descreve uma proposta de uma ferramenta para avaliação de riscos de infecção do Sítio Cirúrgico com a utilização de técnicas de Inteligência Artificial com o propósito de disponibilizar uma ferramenta para que os profissionais da área da saúde façam uso para auxiliar na prevenção de infecções no ambiente cirúrgico.

2 REFERENCIAL TEÓRICO

Neste capítulo, inicialmente é apresentada a Inteligência Artificial e *Machine Learning*, e finalizando a seção é abordada a Inteligência Artificial aplicada à medicina e a importância deste recurso para a área da saúde. Na segunda seção a Infecção Relacionada à Assistência à Saúde é citada, bem como, a Infecção do Sítio Cirúrgico. Na terceira seção é exposta a análise exploratória de dados abordando sobre limpeza de dados e feature engineering. Na seção seguinte é apresentada a estatística descritiva e estatística inferencial e a correlação - medidas de dispersão. Na quinta seção é retratado algoritmos de árvore, os algoritmos de classificação e algoritmos de regressão linear. Na sexta seção é exibido o treinamento e otimização de algoritmos expondo o *Grid Search* e o método *K-Folding*. As métricas de performance de algoritmo são mostrados na sétima seção, incluindo acurácia, recall, *F1 Score*, AUC e ROC curves, matriz de confusão. Por fim, na seção oito são apresentados os trabalhos correlatos referentes ao tema deste trabalho.

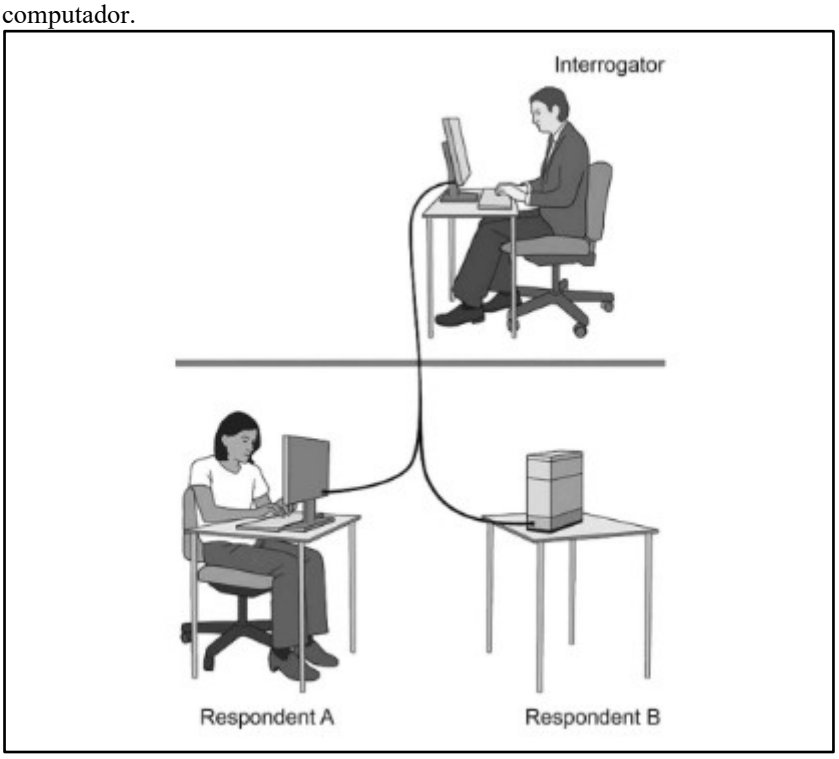
2.1 INTELIGÊNCIA ARTIFICIAL

A inteligência artificial é considerada uma área científica nova, porém sua origem surgiu há 2400 anos quando o filósofo grego Aristóteles idealizou o raciocínio lógico, que resultou na linguagem lógica criada por Leibniz e Newton. Em seguida, a álgebra foi criada por George Boole, introduzindo as bases dos circuitos de computador. O matemático inglês Alan M. Turing que idealizou o conceito de máquina pensante com a concepção do teste de Turing, o qual tinha por objetivo estabelecer se um computador pensa como um humano, anunciando assim o teste de Turing, lançando, desta forma, a inteligência artificial nos anos 1940 e 1950. Vale ressaltar que, John McCarthy mencionou o termo inteligência artificial pela primeira vez em 1956, no *Massachusetts Institute of Technology (MIT*) (GUPTA; NAGPAL, 2020).

Ainda de acordo com Gupta e Nagpal (2020), basicamente, o teste de Turing apresenta três terminais, onde, dois são operados por humanos e um terceiro por meio do computador, uma pessoa é definida como interrogador ou questionador e, outro humano e o computador são identificados são os respondentes, sendo que, o interrogador questiona por meio de um grupo de perguntas e os respondentes humanos e o computador respondem em um tempo pré-determinado sobre uma área específica, assunto e contexto. Neste caso, o

interrogador recebe dois grupos de respostas, no entanto, não tem conhecimento qual se origina do humano e qual é proveniente do computador. Em seguida, por meio de uma avaliação minuciosa, se o interrogador não tiver certeza de quais respostas originou-se do computador, tanto o humano, quanto o computador são considerados respondentes. Sendo assim, o computador passou no teste para comportamento inteligente. No entanto, vale ressaltar que o teste não é tão direto, uma vez que, os seres humanos são superiores aos computadores nos aspectos criatividade, raciocínio e discernimento. Então, as respostas baseadas neste perfil, com certeza o ser humano vencerá, porém, os computadores possuem a vantagem de serem mais precisos, rápidos, principalmente em cálculos. A Figura 1 ilustra o Teste de *Turing*:

Figura 1 - Exemplo do teste de Turing, em que o interrogador deve determinar qual respondente lê o computador.



Fonte: GUPTA; NAGPAL (2020, p.3).

Conforme Russel e Norvig (2010), na *IBM*, *Nathaniel Rochester* e colegas produziram os primeiros programas de IA. Herbert Gelernter (1959) construiu o *Geometry Theorem Prover*, que foi capaz de provar teoremas que muitos estudantes de matemática consideraram complicados.

Na visão de Gupta e Nagpal (2020), sabe-se que os computadores têm habilidades mecânicas para a facilitação de várias práticas através de simples programações onde se

realizam com eficiência e confiabilidade, diferente dos humanos que se mostram entediados com atividades monótonas, já o computador por não raciocinar e não se readaptar, se mostra eficiente nas atividades de monotonia. Ao contrário, o ser humano consegue se adaptar diante de novas situações, por meio de interpretação ao visualizar e extrair dados do que vê, ouvir e o cérebro dar um significado ou sentido ao som. Estudos mostram uma interessante maneira dos humanos solucionar problemas, baseado em pensamento abstrato, raciocínio de alto nível e reconhecimento de padrões.

Na concepção de Russell e Norvig (2013) Inteligência Artificial demonstra-se como uma das ciências mais recentes, com início após a Segunda Guerra Mundial e, atualmente, abrange uma enorme variedade de campos, desde áreas de uso geral, como aprendizado e percepção, até tarefas específicas como jogos de xadrez, demonstração de teoremas matemáticos, criação de poesia e diagnóstico de doenças. Para os autores, a Inteligência Artificial automatiza tarefas intelectuais sendo potencialmente relevante para qualquer esfera da atividade intelectual humana.

Como argumenta Sage (1990), o objetivo da Inteligência Artificial é o desenvolvimento de paradigmas ou algoritmos que requeiram máquinas para realizar tarefas cognitivas, para as quais os humanos atualmente são melhores, pois a IA é capaz de imitar algumas características humanas como percepção visual, reconhecimento de fala, tradução de idiomas, entre outros. Um sistema de IA deve ser capaz de fazer três coisas:

- Armazenar conhecimento;
- Aplicar o conhecimento armazenado para resolver problemas;
- Adquirir novo conhecimento através da experiência.

Além de ser capaz de fazer tais coisas, um sistema de IA deve possuir três componentes fundamentais: representação, raciocínio e aprendizagem. A representação é a maneira que é utilizada para representar o conhecimento genérico sobre um domínio do problema, o raciocínio é a habilidade de resolver problemas, e por fim, a aprendizagem é um processo de mudança de comportamento obtido através da experiência construída por fatores emocionais, neurológicos, relacionais e ambientais (RUSSEL; NORVIG, 2013). Segundo Rouhiainen (2018), IA é a habilidade da máquina de usar algoritmos para aprender e usar o que aprendeu para tomar decisões como um humano, porém com a possibilidade de analisar um massivo número de informações instantaneamente.

Dentro desse contexto temos os agentes inteligentes, que conforme Gupta e Nagpal (2020) consistem em um sistema que percebe seu ambiente, aprende e interagem de forma inteligente, os agentes inteligentes podem ser divididos em duas grandes categorias: agentes de software e físicos.

Sendo agente de *software* o conjunto de programas projetados para realizar tarefas específicas, por exemplo, verificar conteúdo de emails recebidos classificando-os (lixo, importante). Já os agentes físicos apresentam um sistema programável para a realização de tarefas das mais variadas, como por exemplo, robôs usados em indústrias para trabalhos de rotina como montagem, soldagem ou pintura, ou ainda robôs móveis para trabalhos de entrega como distribuição de correios e correspondência em salas diferentes, esses robôs também são usados debaixo d'água para prospectar petróleo (GUPTA; NAGPAL, 2020).

Como caracteriza Nilsson (1998), aprendizado de Máquina então pode ser visto como o processo de mudança de estrutura, programa ou dado, baseado em entradas ou em resposta a informações externas, de maneira que o desempenho futuro de um algoritmo em alguma tarefa melhore. Enquanto Mitchell, Mitchell e Thomas (1997), para afirmar que uma máquina aprende ela deve, a partir de uma experiência E, com respeito a uma classe de tarefas T e medida de performance P, melhorar sua performance na tarefa T, medida por P, com essa experiência E.

Tradicionalmente, métodos de aprendizado de máquina indutivos são classificados em três tipos: aprendizado supervisionado, aprendizado não supervisionado e aprendizado semi supervisionado. A forma e a presença (ou ausência) de supervisão são aspectos primordiais para caracterizar cada uma das três diferentes abordagens (GÉRON, 2019).

De forma resumida, na aprendizagem supervisionada, um conjunto de dados rotulados é fornecido ao algoritmo para que ele aprenda a classificar novos dados com base nesses exemplos fornecidos (*classification*), então, para este tipo de aprendizagem o que se deseja é construir classificadores. Na aprendizagem não supervisionada não é fornecido este conjunto de dados rotulados, portanto o que se deseja é agrupar os dados de acordo com sua similaridade (*clustering),* o algoritmo de aprendizagem não supervisionada procura encontrar padrões nos dados a serem agrupados. Por fim na aprendizagem semi supervisionada também se tem um conjunto de exemplos fornecidos, mas este conjunto é limitado, portanto em uma das abordagens deste tipo de paradigma é a realimentação do conjunto de treinamento com dados classificados pelo próprio algoritmo, as classificações realizadas pelo algoritmo são

acrescentadas ao conjunto de treinamento e o algoritmo é treinado novamente, mas agora com mais dados no conjunto de treinamento (GÉRON, 2019).

2.1.1 *Machine Learning*

Em 1959 o pioneiro em inteligência artificial Arthur Samuel, engenheiro do MIT criou o termo tecnológico "*Machine Learning*" (ML), conceituando como um campo de estudo que dá aos computadores a habilidade de aprender sem terem sido programados. *Machine learning* que em português significa "aprendizado de máquina" se trata de máquinas e sistemas capazes de adquirir novos conhecimentos (SILVA, 2020).

Conforme Géron (2019, p.4), **"**O aprendizado de máquina é a ciência (e a arte) de programar computadores para que eles possam aprender com os dados". Enquanto Samuel (1959, p.211) define que o aprendizado de máquina é o campo de estudo que dá aos computadores a capacidade de aprender sem serem explicitamente programados.

Bernard (2017) aponta que Arthur Samuel, o pioneiro da Inteligência artificial, trabalhava na época, em um projeto para criar uma máquina autônoma com tais características, mas somente com um advento da internet que o machine learning começou a tomar forma, pois com tanta informação coletada e armazenada na web se fez necessário criar meios de organizar todo o conteúdo de forma automatizada, tendo aí um dos pilares do machine learning, pois tem a função de analisar dados e detectar padrões.

Como descrito por Rabelo (2018), o *machine learning* atua por meio de algoritmos e *big data*, identificando padrões de dados e criando conexões entre eles para aprender a executar uma tarefa sem a ajuda humana e de forma inteligente. Esses algoritmos usam análises estatísticas para prever respostas mais precisamente e entregam o melhor resultado preditivo com menos chance de erro.

Conforme Géron (2019) os sistemas de aprendizado de máquina podem ser classificados de acordo com a quantidade e o tipo de supervisão que recebem durante o treinamento. Existem quatro categorias principais: aprendizagem supervisionada, aprendizagem não supervisionada, aprendizagem semi-supervisionada e aprendizagem por reforço.

Nesta mesma linha de considerações, Coelho (2020) descreve estas quatro categorias de sistemas de aprendizado de máquina:

- **Aprendizado supervisionado**: São exemplos rotulados onde o algoritmo de aprendizagem recebe um conjunto de entradas junto com as saídas corretas correspondentes, e o algoritmo aprende comparando a saída real com as saídas corretas para encontrar erros, logo, ele modifica o modelo de acordo;
- **Aprendizado não supervisionado**: é usado contra dados que não possuem rótulos, ou seja, o sistema não sabe a "resposta certa". O algoritmo deve descobrir o que está sendo mostrado e o objetivo é explorar os dados e encontrar alguma estrutura neles, sendo assim, o aprendizado não supervisionado funciona bem em dados transacionais;
- **Aprendizado semi supervisionado**: frequentemente usado nas mesmas aplicações que o aprendizado supervisionado, porém ele pode usar tanto dados rotulados quanto não marcados para o treinamento. Esse tipo de aprendizagem pode ser usado com métodos como a classificação, regressão e previsão. É bastante útil quando o custo associado à rotulagem é alto para permitir um processo de treinamento totalmente rotulado, como exemplo, a identificação do rosto de uma pessoa em uma *webcam*;
- **Aprendizado por reforço:** muito usado para a robótica, jogos e navegação, onde o algoritmo descobre por meio de tentativa e erro quais ações geram as melhores recompensas.

De acordo com, Silver Shark Solutions (2018, n.p),

> O *machine learning* nada mais é do que um subconjunto da IA, ou seja, é uma das formas de utilizar a Inteligência artificial onde toda a aprendizagem de máquina conta com a IA, mas nem toda IA tem aprendizado automático. [...] o machine learning tem a capacidade de modificar-se quando exposto a mais dados, ou seja, o aprendizado automático da máquina é dinâmico e não requer intervenção humana para realizar certas mudanças, tornando-o menos frágil e menos dependente de especialistas humanos.

Certamente o *machine learning* é uma tecnologia que vem evoluindo, nasceu do reconhecimento de padrões e da teoria de que os computadores poderiam aprender sem ser programados para executar tarefas específicas. O aspecto interativo da aprendizagem de máquina é importante, porque à medida que os modelos são expostos a novos dados, eles são capazes de se adaptar independentemente, assim, eles aprendem com cálculos anteriores para produzir decisões e resultados confiáveis e repetíveis (SAS, 2021).

SAS (2021) ainda completa que, mesmo o *machine learning* tendo algoritmos criados a muito tempo, sua capacidade de aplicar automaticamente cálculos matemáticos complexos de forma cada vez mais rápida é um desenvolvimento recente.

Conforme enfatiza Rabelo (2018), na opinião de alguns especialistas da área, presumi-se que o *machine learning* prosseguirá evoluindo em todo o espaço do mercado móvel, dentro de aplicativos, assistentes digitais e IA como um todo, podendo entrar no território dos drones e auto-condução dos carros. Contudo, como a demanda por mais dados e mais algoritmos está aumentando espera-se que mais ferramentas de *machine learning* se tornem disponíveis.

Conforme Géron (2019), ao refirir-se a "Aprendizado de Máquina", logo se imagina um robô: um mordomo confiável. No entanto, o aprendizado de máquina não é apenas uma fantasia futurística, ela está presente, pois já existe há décadas em algumas aplicações especializadas, como o *Optical Character Recognition* (*OCR),* em português Reconhecimento Óptico de Caracteres. O primeiro aplicativo de *Machine Learning* que realmente se tornou popular conquistando o mundo na década de 1990 foi o filtro de *spam*. Não é exatamente uma *Skynet* autoconsciente, mas tecnicamente se qualifica como Aprendizado de Máquina, isto foi seguido por centenas de aplicativos de ML que agora alimentam silenciosamente centenas de produtos e recursos que você usa regularmente, desde melhores recomendações até pesquisa por voz.

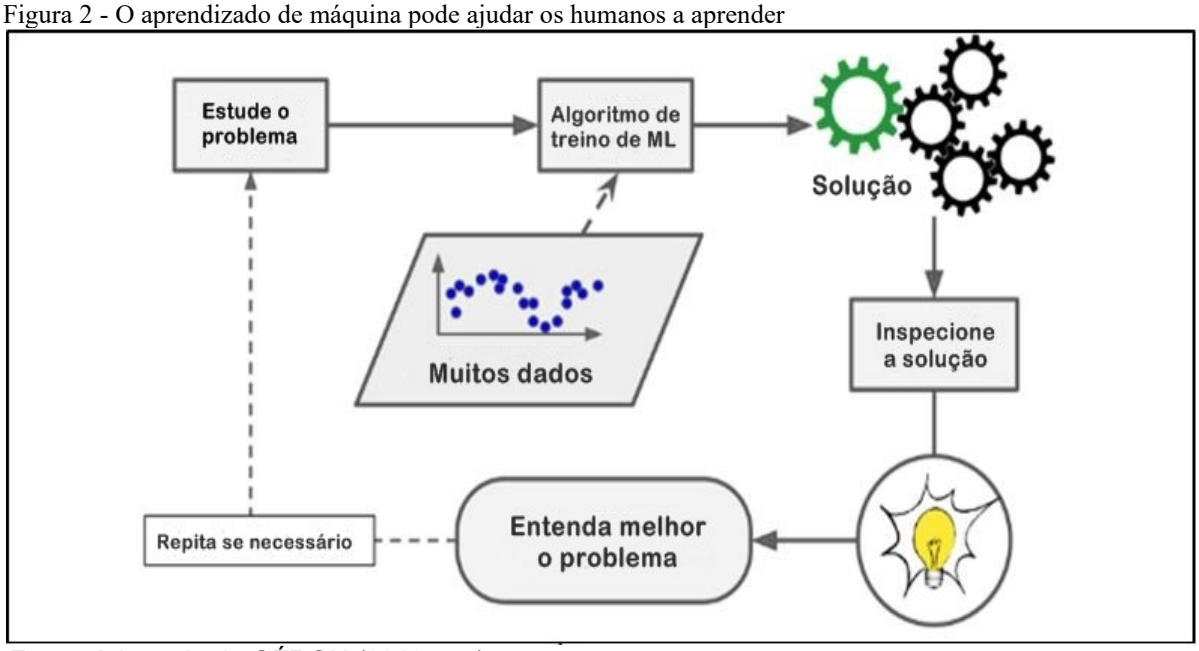
Diante de tantos apontamentos fica claro que a tecnologia *machine learning* apresenta inúmeras vantagens, pois tem capacidades de processar de forma ilimitada, dados das mais diversas fontes, podendo revisar e ajustar com base no comportamento do cliente, sendo identificadas as variáveis para transmitir informações de forma certa podendo ainda automatizar processos. Outro ponto a ser observado é a velocidade com que ML consome dados e identifica informações em tempo real, esses sistemas atuam sobre os resultados do aprendizado da máquina e tornam a mensagem de marketing muito mais dinâmica, ajudando na retenção e conversão para processar informações de forma rápida. Têm ainda como grande vantagem os modelos que podem aprender com resultados passados para melhorar continuamente suas previsões com base em dados novos. O machine learning pode também ser usado para identificar vários segmentos, bem como criar micro-segmentos com base em padrões comportamentais (DASHGOO, 2018).

Por fim, e não menos importante o *Data Mining*, que em português se chama mineração de dados, tem por função aplicar técnicas de ML para pesquisar grandes

quantidades de dados para poder ajudar a descobrir padrões que não eram imediatamente aparentes (DASHGOO, 2018).

A Figura 2 ilustra como o aprendizado de máquina pode contribuir para o aprendizado dos seres humanos.

Figura 2 - O aprendizado de máquina pode ajudar os humanos a aprender



Fonte: Adaptada de GÉRON (2019, p.5).

De forma resumida, pode-se concluir que o aprendizado de máquina atua na resolução de problemas que exigem ajustes finos como um algoritmo que podem muitas vezes simplificar o código e ter um desempenho melhor do que a abordagem tradicional atual também em problemas complexos para os quais o uso de uma abordagem tradicional não produz uma boa solução, tem ainda ação em ambientes flutuantes, no qual um sistema de aprendizado de máquina pode se adaptar a novos dados e por fim obter insights sobre problemas complexos e grandes quantidades de dados.

2.1.2 Inteligência Artificial aplicada a medicina

Inteligência Artificial em medicina é o uso de computadores que, analisando um grande volume de dados e seguindo algoritmos definidos por especialistas na matéria, são capazes de propor soluções para problemas médicos (LOBO, 2017).

Ainda segundo Lobo (2017) a relação da inteligência artificial e a medicina, já vêm de algumas décadas, onde uma das motivações deu-se pela capacidade do computador em

armazenar grandes volumes de dados e a velocidade em capturar estes dados, criou-se a expectativa de que o especialista da área de saúde pode ser beneficiado com um instrumento que melhorasse a tomada de decisão em relação ao diagnóstico e ao trato do paciente.

Com a associação das duas áreas de conhecimento, percebe-se a dificuldade em elaborar software com detalhamento e precisão para o auxílio à tomada de decisão, boa parte desta dificuldade se deu pela falta de padronização dos dados e a diversidade na inferência realizada pelo especialista ao elaborar o diagnóstico (GÓMEZ-GONZÁLEZ *et al.*, 2020).

No contexto de um hospital, conforme enfatizam Gómez-González *et al.* (2020), o processo de diagnóstico que o especialista da área de saúde realiza, pode estar diretamente relacionado à qualidade dos dados disponíveis e o nível de conhecimento na especialidade em questão. Sendo assim, os dados do prontuário do paciente devem ser de quantidade e qualidade suficientes para a identificação do diagnóstico e a conduta de tratamento.

Ainda nesta mesma de considerações Gómez-González *et al.* (2020), descrevem que em um ambiente hospitalar, o diagnóstico de um especialista da área de saúde, pode estar diretamente relacionado à qualidade dos dados disponíveis, a experiência profissional e o nível de conhecimento do profissional. Sendo assim, os dados do prontuário devem ter quantidade e qualidade suficiente para um bom diagnóstico e conduta de tratamento. A padronização e o preenchimento correto dos dados do paciente são fatores fundamentais para um bom diagnóstico. O especialista da área de saúde deve ter condições de selecionar adequadamente os dados e saber quando a experiência pessoal deve ser empregada.

Em um trabalho recente, Mukherjee (2017) relata que foi desenvolvida na Universidade de Columbia: uma radiologista discutia a importância de diagnosticar precocemente um Acidente Vascular Cerebral (AVC) numa tomografia computadorizada, permitindo a destruição oportuna de um coágulo no cérebro. É fácil diagnosticar um AVC quando o cérebro já está morto e cinzento, dizia ela. O desafio é diagnosticar e intervir precocemente. O reconhecimento de imagens feito por IA poderia obviar esse desafio pelo reconhecimento de pequenos detalhes indicando áreas suspeitas em cortes de CT que poderiam passar despercebidas.

Conforme Gómez-González *et al.* (2020) argumentam, os números globais e o mercado de IA em Medicina e Saúde preveem impactos positivos muito relevantes para os próximos anos. No entanto, a análise econômica deve incluir os pontos éticos e sociais relacionados aos sistemas de saúde, indústrias e pacientes. É importante notar que o custo de decodificação de um genoma humano é substancialmente baixo na ordem de algumas

centenas de euros, mas os preços de alguns dos tratamentos mediados por IA, como certos medicamentos personalizados, podem atingir valores "impossíveis", mesmo na ordem de milhões de euros por caixa. Essa etapa íngreme se deve às dificuldades de adequar individualmente as moléculas do medicamento ao genoma específico de um indivíduo. Novos modelos de cobertura de saúde, seguro e acessibilidade podem ser necessários, como tal, tecnologias clinicamente excelentes representam um risco claro de evoluir para um fator crescente de desigualdade para a maioria das pessoas.

Gómez-González *et al.* (2020) chamam a atenção para o duplo uso das tecnologias quando se referem aos aspectos e aplicações das tecnologias mediadas por IA, uma vez que, argumentam que estas podem ser muito questionáveis e que, mesmo que a sociedade não dê atenção neste momento a estas particularidades, as mesmas continuam em constante evolução de modo não controlado e não supervisionado. Mesmo assim, os autores destacam que, grande parte dos desenvolvimentos demonstra ao mesmo tempo potencial de melhorias muito favoráveis e perturbadoras, gerando resultados altamente impressionantes e eticamente controversos, e igualmente muito negativos.

De acordo com Gómez-González *et al.* (2020), a neurociência é uma das áreas científicas em que os avanços tecnológicos estão promovendo um desenvolvimento extraordinário. As principais contribuições vêm da fusão de IA, fotônica, que é a ciência da geração e detecção de luz, e engenharia. A neurociência também abrange neurocirurgia e neurologia com outras áreas clínicas como farmacologia, psicologia, ciências, biologia e genética e bioquímica. Tendo as questões éticas como desafio, os avanços apontam para a leitura e decodificação remota progressivamente não invasiva de sinais complexos do cérebro e o design de interfaces homem-máquina avançadas.

Além disso, em meio a tantos benefícios, todo o conhecimento através da IA abre caminho para o sistema nervoso humano já começando a permitir o controle interativo de próteses ativas e inovadoras, que oferecem uma grande esperança para muitas pessoas com condições severamente incapacitantes e muitas aplicações industriais potenciais. Existem projetos fortemente financiados, por exemplo, o *European Human Brain Project* e, a *USA Brain Initiative*, orientado para “os aspectos positivos” da neurociência. No entanto, esse conhecimento também está relacionado a caminhos muito controversos, por exemplo, a capacidade potencial de "ler a mente" e sua eventual combinação com diferentes tipos de estimulação neural e interação com sinais cerebrais, que, por sua vez, pode levar a formas indesejadas de manipulação (falta de livre arbítrio) e controle humano. Um exemplo claro de

uma área controversa mediada por IA é a edição de genes humanos. A possibilidade de alterar e substituir fragmentos definidos da cadeia genética (ácido desoxirribonucleico, DNA) de células humanas é uma tarefa desafiadora que requer ferramentas de IA para atingir a precisão necessária - com muitas aplicações potenciais (GÓMEZ-GONZÁLEZ *et al.*, 2020).

Ainda nesta mesma linha de considerações, Gómez-González *et al*. (2020), apontam as seguintes indagações, é ético projetar “seres humanos experimentais”? O que acontecerá se essas pessoas se tornarem adultas e procriarem? Haverá algum limite nos genes a serem modificados? Ou seja, se o procedimento for tecnicamente viável, alguém projetaria alguns tipos de “super-humanos”? Deve haver algum tipo de supervisão desta pesquisa? A maioria dessas questões é comum a outras áreas controversas das aplicações de IA, mal foram feitas e ainda não têm uma resposta clara.

Gómez-González *et al*. (2020) enfatizam que, um dos avanços a partir da aplicação da IA é dado pelo desenvolvimento recentemente divulgado de um protótipo de uma 'máquina viva' - também considerado um 'robô biológico' ('*biobot*') baseado em células animais e a busca por formas artificiais de vida para aplicações militares. Com um laboratório relativamente pequeno, conhecimento e as ferramentas adequadas, a edição de genes pode ser alcançada. Existem alguns relatos de pessoas em auto experimentação, que modificam seu DNA e o reintroduzem em seus corpos, em um novo tipo de “abordagem da ciência cidadã” chamada *biohacking*. Sua finalidade geral é obter "capacidades aprimoradas", mas este tipo de pesquisa também apresenta um interesse potencial para o mercado de saúde.

De acordo com Forato (2020), no Brasil, algoritmos desenvolvidos com IA já são utilizados em hospitais e laboratórios brasileiros, como dentro do Fleury Medicina e Saúde. Na área de exames de imagem, o laboratório validou um algoritmo que auxilia os médicos no diagnóstico de embolia pulmonar — doença em que artérias pulmonares são bloqueadas por um coágulo sanguíneo, que pode levar ao óbito.

No entanto, a disponibilidade de certas ferramentas mediadas por IA (*hardware, software* e conjuntos de dados) também pode abrir o caminho para formas maliciosas de *biohacking*. A edição de genes combinada com o uso de tecnologias de IA apresenta um limite preocupante e muito suave para o armamento e o bioterrorismo. Organismos geneticamente modificados podem ser muito difíceis de detectar e rastrear. Eles podem até mesmo ser armazenados e transportados como células alteradas em uma pessoa portadora visando indivíduos ou populações específicas (GÓMEZ-GONZÁLEZ *et al.*, 2020).

É importante destacar ainda, a proposta de classificação de impacto social da IA proposta por Gómez-González *et al.* (2020), considerando seu trabalho com diagnósticos apoiados por IA estão entre os mais acessíveis, no entanto implementáveis, mas trazem benefício social e no Anexo A, apresenta-se uma classificação graduada de IA e aplicações mediadas por IA em Medicina e Saúde de acordo com seu caráter benéfico versos prejudiciais, conforme reconhecido na literatura revisada. Para cada uma dessas aplicações, relatam a tecnologia, as implementações específicas por trás e seu nível de disponibilidade de acordo com referências publicadas. De um lado do espectro do impacto social, consideram os aplicativos que se mostraram benéficos para apoiar a tomada de decisões clínicas, enquanto do outro lado colocam os aplicativos que são amplamente considerados prejudiciais. Propõe-se ainda uma (nova) escala denominada "*Technology Availability Level*" (TAL) que dá uma descrição qualitativa do grau de disponibilidade de uma tecnologia, em uma escala numérica de 0 (desconhecido) a 9 (disponível para o geral público). A escala TAL é semelhante em formato (e relacionada) ao padrão "Níveis de preparação de tecnologia" (TRL), mas, como mencionado, é baseada em referências publicadas (na literatura científica e acadêmica, relatórios industriais ou corporativos e fontes de citações da mídia em geral considerado confiável de acordo com os padrões). É importante considerar que "disponibilidade" não é necessariamente equivalente a "níveis de prontidão" devido a fatores como divulgação de acordo com estratégias industriais, proprietárias e / ou governamentais, e que a escala TAL também não avalia o cumprimento dos processos regulatórios.

Os valores definidos para a escala TAL são os seguintes (GÓMEZ-GONZÁLEZ *et al.*, 2020):

TAL 0. Status desconhecido. Não considerado viável de acordo com as referências.

TAL 1. Status desconhecido. Considerado viável de acordo com referências indiretas relacionadas. TAL 2. Ideia geral/básica proposta publicamente.

TAL 3. Abertura de convites para financiamento público de I&D.

TAL 4. Divulgação de resultados de projetos acadêmicos/parciais. TAL 5.

Divulgação do design inicial do produto.

TAL 6. Protótipo operacional/'primeiro caso' divulgado. TAL 7.

Produtos divulgados, mas não disponíveis.

TAL 8. Disponível para usuários restritos (por exemplo, profissionais). TAL 9.

2.2 INFECÇÃO RELACIONADA À ASSISTÊNCIA À SAÚDE – IRAS

As Infecções Relacionadas à Assistência à Saúde (IRAS) são definidas como toda infecção que ocorre no paciente durante o processo de atendimento em algum estabelecimento de saúde, no período de 48 a 72 horas do primeiro contato com o sistema de saúde. As taxas de infecções são maiores em países em desenvolvimento e em unidades de terapia intensiva (FERRAZ *et al.*, 2019).

Prates afirma que as Infecções Relacionadas à Assistência à Saúde são reconhecidas mundialmente como um grave problema de saúde pública por serem os eventos adversos mais frequentes relacionados à assistência ao paciente e estarem associadas a uma alta morbimortalidade, aumento do tempo de permanência e dos custos hospitalares (PRATES *et al.*, 2018).

As infecções hospitalares e as preocupações com os microrganismos encontrados nos pacientes hospitalizados foram documentadas inicialmente por especialistas da área da saúde como *Ignaz Semmelweis*, *Florence Nightingale* e *Joseph Lister* em meados do século XIX, durante a chamada "revolução pasteuriana" (LARSON, 1989).

Entre as IRAS, a Infecção de Sítio Cirúrgico tem papel de destaque podendo manifestar-se até 30 dias após a cirurgia ou em até 90 dias, se houver colocação de implantes sendo classificadas de acordo com o grau de acometimento, a saber: incisional superficial, incisional profunda ou infecção de órgão e/ou cavidade (PRATES *et al*., 2018).

No Brasil, na década de 1990 foram publicados diversos documentos que relatavam as preocupações com as infecções hospitalares e desde então este assunto vem sendo debatido. Considera-se uma infecção hospitalar toda a infecção que é adquirida durante uma internação ou até mesmo após a alta do paciente, quando é possível relacionar com a internação do cliente (BRASIL, 2017).

Nos dias de hoje, não é mais utilizado o termo infecções hospitalar, e sim, infecções relacionadas à assistência à saúde, pois assim, se pode envolver todas as infecções que são relacionadas à assistência do paciente em qualquer ambiente, seja ele hospitalar ou não, como por exemplo, as clínicas médicas que realizam pequeno procedimentos. (HORAN; ANDRUS; DUDECK, 2008). Podendo ocorrer infecções relacionadas ao trato urinário, na corrente sanguínea, nas incisões cirúrgicas e ao trato respiratório (BRASIL, 1997; BRASIL, 2013).

No Brasil, o impacto das IRAS sobre as taxas de morbidade, mortalidade e sobre os custos da assistência hospitalar prestada foi reconhecido apenas em 1983, quando a portaria

Nº 196 foi criada, estabelecendo que todo hospital dever constituir uma comissão de controle de infecções hospitalares (MS, 1983).

Dessa forma, a escassez de dados nacionais relacionados às infecções hospitalares consiste em um entrave para o desenvolvimento de estratégias de controle e prevenção das IRAS. Taxas significativas de IRAS e os agravos decorrentes destas resultam em elevados custos para os serviços de saúde públicos e privados. Cerca de dois milhões de casos e 80.000 mortes por ano nos EUA apresentam custo estimado entre 4,5 e 5,7 bilhões de dólares. As IRAS contribuem de forma significativa para a permanência dos pacientes por longos períodos, elevação dos custos hospitalares e também na letalidade, sendo que os países em desenvolvimento podem ter até 20 vezes mais infecções que os países desenvolvidos (PITTET *et al.*, 2008).

O aumento dos índices de infecções nos países desenvolvidos se dá pela falta de qualificação profissional, diminuição de efetivo, junto com a má estrutura física e o desconhecimento das boas práticas que auxiliam na redução das infecções (PADOVEZE; FORTALEZA, 2014).

Conforme CDC (2014) enfatiza atualmente a saúde emprega muitos tipos de dispositivos e procedimentos invasivos para tratar pacientes e ajudá-los a se recuperar, onde muito provavelmente tais infecções estão associadas aos dispositivos usados em procedimentos médicos, como cateteres ou ventiladores. Essas infecções associadas aos cuidados de saúde (HAIs) incluem infecções da corrente sanguínea associadas ao cateter, infecções do trato urinário associadas ao cateter e pneumonia associada ao ventilador. Nesse sentido Brasil (2007), as infecções também podem ocorrer nos locais da cirurgia conhecida como infecções do local da cirurgia. As infecções hospitalares geram uma grande preocupação para os nosocômios, principalmente as IRAS que são ocasionadas por microrganismos resistentes a mais de duas classes de antibióticos,

Como faz notar Who (2005), essa resistência aos antibióticos é um problema que atinge todo o mundo, e ocorre devido a inúmeras causas, como por exemplo, o uso indiscriminado de antimicrobianos, a inexistência de programas que façam o controle de infecção da forma correta, além de laboratórios que não prestam suporte adequado.

Para que possa ser evitado o surgimento das infecções é necessário realizar a vigilância epidemiológica das IRAS. Tal ação visa gerar informações que serão utilizadas para promover a melhoria do ambiente, além de auxiliar na criação de estratégias de

prevenção e controle de infecções. Além de que, estar ciente das informações que tem relação com as IRAS, faz com que sejam prevenidos possíveis surtos (BRASIL, 1997).

O *Centers for Disease Control and Prevetion* (CDC) Centro de Controle e Prevenção de Doenças é uma agência do Departamento de Saúde e Serviços Humanos dos Estados Unidos, na qual trabalha para monitorar e prevenir essas infecções porque são uma ameaça importante para a segurança do paciente (CDC, 2014).

A vigilância epidemiológica das IRAS deve ser planejada e realizada em conjunto com a Comissão de Controle de Infecção Hospitalar (CCIH). Pois esta é uma equipe treinada e que tem conhecimento dos conceitos epidemiológicos. Todas as estratégias/treinamentos preferencialmente têm que ser multimodais, ou seja, englobando vários métodos, como treinamentos realísticos, prática a beira leito, indicadores, entre outros. Além disso, a higiene de mãos deve fazer parte de todos os treinamentos educativos de prevenção de IRAS, devendo ser abordado à técnica correta e os momentos adequados (BRASIL, 1997).

2.2.1 Infecção do Sítio Cirúrgico

As Infecções de Sítio Cirúrgico (ISC) estão entre as mais prevalentes nas instituições de saúde. No ano de 2011 nos EUA, as ISC acometeram em média 157.500 mil pacientes. O risco de morte atribuível a este tipo de infecção é alto, variando de 33 a 77%, sendo associado a um aumento de 2 a 11 vezes para o desfecho de óbito (FERRAZ *et al.*, 2019).

Conforme Brasil (2013), as infecções de sítio cirúrgico são classificadas em:

ISC Incisional Superficial: São as infecções que ocorrem nos primeiros 30 dias após a cirurgia, mas vai envolver apenas a pele e o subcutâneo do paciente. Deve conter pelo menos um destes agravantes:

- Drenagem purulenta no local da incisão superficial;
- Cultura positiva que foi coletada de forma asséptica e não por swabs;
- Reabordagem cirúrgica, se observado: dor, sensibilidade, edema e/ou hiperemia;
- Diagnóstico pelo médico assistente.

ISC Órgão/Cavidade: Ocorre desde os primeiros 30 dias até um ano após a cirurgia se houver a colocação de próteses, além de envolver qualquer órgão ou cavidade que tenha sido manipulada durante o procedimento. Deve conter pelo menos um destes agravantes:

- Cultura positiva que foi coletada de forma asséptica e não por swabs;

- Abscesso;
- Diagnóstico pelo médico assistente.

ISC Incisional Profunda: Pode ocorrer desde os primeiros 30 dias até um ano após a cirurgia se houver a colocação de próteses, além de envolver os tecidos moles profundos à incisão, como por exemplo, a fáscia ou músculos. Deve conter pelo menos um destes agravantes:

- Drenagem purulenta no local da incisão profunda;
- Deiscência de pontos parcial ou total;
- Reabordagem cirúrgica, se observado: temperatura axilar >38ºC, dor e/ou sensibilidade;
- Abscesso;
- Diagnóstico pelo médico assistente.

Nos últimos anos, é possível verificar que houve um elevado índice de procedimentos cirúrgicos. Como cirurgias ortopédicas, cardiovasculares e retiradas de tumores, sendo que isso se deve ao aumento da expectativa de vida e também da violência. (BRASIL, 1997).

Segundo as estimativas, são realizadas anualmente até 281 milhões de cirurgias de grande porte, ou seja, uma cirurgia para cada 25 pessoas. Apesar da evolução e da melhoria da técnica cirúrgica, as complicações ainda aparecem e podem chegar a 16%, sendo que os óbitos em países em desenvolvimento podem alcançar a marca de 10% (BRASIL, 1997).

Quando se analisa todas as complicações que podem ocorrer nas cirurgias, nos deparamos com as Infecções relacionadas à assistência à saúde (IRAS), sendo que no pós-operatório está é a complicação mais comum, ocorrendo em até 20% dos procedimentos realizados, podendo levar a um aumento significativo da morbidade e da mortalidade do cliente. No Brasil, observamos que as Infecções do Sítio Cirúrgico estão ocupando o terceiro lugar, ocorrendo entre 14 a 16% dos pacientes hospitalizados (BRASIL, 1997).

Para Prates *et al.* (2018) a ISC é um dos principais alvos da vigilância epidemiológica nas instituições de saúde. Nos países subdesenvolvidos e em desenvolvimento, estima-se que a ISC possa acometer até um terço dos pacientes submetidos a procedimentos cirúrgicos e, embora menos frequente nos países industrializados, ocupa o segundo lugar dentre as IRAS na Europa e nos Estados Unidos, sendo que neste último,

acomete 2 a 5% dos pacientes, totalizando 160.000 a 300.000 episódios anualmente. No Brasil a ISC é considerada um dos principais riscos relacionados à segurança do paciente nos serviços de saúde e dentre todas as IRAS, ocupa a terceira posição, compreendendo 14 a 16% daquelas identificadas em pacientes hospitalizados.

A ISC leva a graves consequências, incluindo o aumento nos gastos devido ao seu tratamento e a um aumento do tempo de internação. O risco de morte dos pacientes com ISC mostra-se aumentado quando comparado aos que não desenvolveram a infecção (CARVALHO *et al.*, 2017).

As Infecções de Sítio Cirúrgico podem causar inúmeros prejuízos para os pacientes, não somente físicos, mas também psicológicos e financeiros, podendo levar a um aumento da permanência deste cliente no hospital, elevar as chances de uma nova abordagem cirúrgica e de reinternações frequentes. Por isso, quando se leva em consideração todos esses aspectos, observa-se que as ISC são as IRAS de maior custo, porém, assim como todas as infecções hospitalares é possível que se previna adotando os cuidados sugeridos pelos guias, manuais, protocolos e boas práticas (BRASIL, 1997).

De acordo com Casali *et al.* (2019), a ISC é a mais comum em usuárias hospitalizadas, superando apenas as Infecções do Trato Urinário (ITU). Quando uma mulher desenvolve ISC pós-cesariana tem cinco vezes mais probabilidade de retornar ao serviço de saúde em pelo menos 30 dias pós-cirurgia; duas vezes mais chances de morrer e exigem em média um adicional de insumos na instituição de três milhões de reais para seu cuidado e tratamento.

Para o Centro de Controle e Prevenção de Doenças (CDC) (2010) a infecção do sítio cirúrgico é uma infecção que ocorre após a cirurgia na parte do corpo onde a cirurgia foi realizada. As infecções do sítio cirúrgico, às vezes, podem ser infecções superficiais envolvendo apenas a pele. Outras infecções do sítio cirúrgico são mais graves e podem envolver tecidos sob a pele, órgãos ou material implantado. O CDC fornece diretrizes e ferramentas para a comunidade de saúde para ajudar a acabar com infecções de sítio cirúrgico e recursos para ajudar o público a entender essas infecções e tomar medidas para proteger sua própria saúde, quando possível.

É indispensável que se adote medidas de boas práticas, que auxiliem na prevenção das ISC. Sendo elas: antibioticoprofilaxia, tricotomia, controle da glicemia no pré e pós-operatório, controle da temperatura durante todo o procedimento cirúrgico, utilizar produtos à base de álcool para preparação da pele, controlar o aporte de oxigênio, aplicar listas de

verificações de segurança, além de educar a família e o paciente para que tenham os corretos cuidados em casa (BRASIL, 1997).

Ainda nesta mesma linha de considerações Scardoni *et al.* (2020) enfatizam que as Infecções Associadas à Assistência à Saúde (IRAS) são os eventos adversos mais frequentes na área da saúde e uma preocupação de saúde pública global. A vigilância é a base para uma prevenção e controle eficazes de *Healthcare-Associated Infections* (HAIs). A vigilância manual exige muito trabalho, é cara e carece de padronização. A inteligência artificial e o *Machine Learning* podem apoiar o desenvolvimento de algoritmos de vigilância de HAI com o objetivo de compreender os fatores de risco de HAIs, melhorar a estratificação de risco do paciente, identificação de vias de transmissão, detecção oportuna ou em tempo real. Há poucas evidências disponíveis sobre a implementação de IA e ML no campo de HAIs e nenhum padrão claro surge sobre seu impacto.

Ainda, conforme Scardoni *et al.* (2020), em estudo, são documentados os potenciais resultados de usar IA para prever infecções:

- Significado para a prática clínica e saúde pública;
- Em um futuro próximo, os sistemas de controle de HAIs baseados em ML podem ser integrados na prática clínica hospitalar;
- Os sistemas de controle HAIs baseados em ML no futuro podem melhorar a eficácia e reduzir os custos das intervenções de segurança do paciente;
- O ML pode levar a uma melhor compreensão dos fatores de risco de HAIs, melhor estratificação de risco do paciente, bem como HAIs oportunas ou em tempo real;
- Detecção e controle;
- Para a implementação e uso de sistemas de controle de HAIs baseados em ML na prática clínica, grandes volumes de dados eletrônicos de saúde devem estar disponíveis, acessíveis e vinculáveis;
- A colaboração fortalecida e multidisciplinar entre as disciplinas de TI e clínicas está prevista para promover a adoção e uso de sistemas de controle HAIs baseados em ML na prática clínica.

Também de acordo com o artigo de Scardoni *et al.* (2020), a revisão identificou 27 estudos nos quais modelos baseados em ML foram aplicados para vigilância e controle de HAIs em diferentes ambientes clínicos. Em geral, há evidências moderadas de que os modelos baseados em ML têm desempenho igual, ou melhor, em comparação com abordagens não ML

e que alcançam padrões de desempenho relativamente altos. No entanto, a heterogeneidade entre os estudos foi muito alta e não se dissipou significativamente nas análises de subgrupos, por tipo de infecção ou tipo de resultado. Mais da metade dos estudos foram conduzidos nos Estados Unidos e a maioria dos estudos focou em infecções de sítio cirúrgico. Os dados comparativos disponíveis são entre diferentes modelos baseados em ML e: pontuações clínicas, modelos de vigilância padrão ou automatizados (baseados em regras) e algoritmos estatísticos de regressão logística; 63% dos estudos tiveram uma abordagem preditiva, enquanto 37% tiveram uma abordagem retrospectiva para a identificação de riscos para HAIs.

Para Fitzpatrick, Doherty e Lacey (2020) um dos principais desafios da IA na vigilância de HAI é a exigência de um conjunto de dados representativo de alta qualidade para desenvolver modelos precisos para cada contexto em que são usados. Uma revisão sistemática de aprendizado de máquina em cuidados intensivos observou que muitos estudos usam conjuntos de dados que são muito pequenos para avaliar o potencial total das aplicações de IA e diretrizes sobre metodologia e validação de previsões são necessárias para ajudar a traduzir os resultados na prática clínica diária. Além do tamanho e da integridade, bancos de dados pré-existentes podem estar usando a Inteligência Artificial na Prevenção de Infecções.

Fitzpatrick, Doherty e Lacey (2020) ainda realçam que, é inerentemente enviesado pela prática clínica e pela prestação de cuidados de saúde da época, o que pode comprometer o atendimento ao paciente se esses preconceitos forem incorporados a um modelo de aprendizado de máquina.

Como acontece com qualquer publicação ou diretriz, a generalização dos modelos de aprendizado de máquina para vigilância de HAI desenvolvidos a partir de dados em um único ambiente de saúde provavelmente será limitada. Por exemplo, em um modelo de CDI, os fatores de risco para CDI em um ambiente eram protetores em outra instituição, o que pode refletir vieses locais ou diferenças nas vias de CDI (FITZPATRICK; DOHERTY; LACEY, 2020).

2.3 ANÁLISE EXPLORATÓRIA DE DADOS

Como apontam Raschka e Mirjalili (2017) a Análise Exploratória de Dados ou *Exploratory Data Analysis* (EDA) é uma primeira etapa importante e recomendada antes do treinamento de um modelo de aprendizado de máquina. Antigamente chamada apenas de Estatística Descritiva, a Análise Exploratória de Dados constitui o que a maioria das pessoas

entende como Estatística, e inconscientemente usam no dia a dia. Consiste em resumir e organizar os dados coletados através de tabelas, gráficos ou medidas numéricas, e a partir dos dados já resumidos, procurar alguma regularidade ou padrão nas observações, o que envolve interpretar os dados, a partir dessa interpretação inicial é possível identificar se os dados seguem algum modelo conhecido que permita estudar o fenômeno sob análise, ou se é necessário sugerir um novo modelo.

Com início em 1977 com *J.W. Tukey*, a análise exploratória visa aumentar o conhecimento do pesquisador sobre uma população a partir de uma amostra. Dessa forma, podemos descrever a (EDA) como um conjunto de métodos adequados para a coleta, exploração, descrição e interpretação de conjuntos de dados numéricos. Tais métodos permitem a exploração dos dados com intuito de identificar padrões de interesse e a representação dos dados caracterizados por estes padrões. Embora as técnicas da (EDA) sejam simples, geralmente são métodos robustos (válidos para uma grande gama de situações e modelos) e resistentes (insensíveis aos erros grosseiros ou dados estranhos) (LOPES *et al.*, 2019).

Não é exagero reafirmar a citação de Tukey (1977, p.1): “A EDA é trabalho de detetive, procurando pistas e evidências; análise confirmatória de dados é trabalho judicial ou quase judicial, que analisa, avalia e julga as provas e as evidências”.

Na visão de Peck, Olsen e Devore (2016) o processo de análise de dados pode ser visto como uma sequência de etapas que levam do planejamento à coleta de dados e a tomada de conclusões fundamentadas com base nos dados resultantes.

Ainda segundo Peck, Olsen e Devore (2016), o processo de análise de dado pode ser organizado e definido em seis etapas:

- Compreender a natureza do problema, afinal analisar dados de forma eficaz requer uma compreensão do problema de pesquisa, onde se deve saber o objetivo da pesquisa e que perguntas espera-se responder. É importante ter uma orientação clara antes de coletar os dados, para garantir a capacidade de responder às perguntas de interesse usando os dados coletados;
- Decidir o que medir e como medir, sendo que esta etapa do processo é decidir quais informações são necessárias para responder as perguntas de interesse. Em alguns casos, a escolha é óbvia. Por exemplo, em um estudo sobre a relação entre o peso de um jogador de futebol da Divisão I e a posição jogada, você precisaria coletar dados sobre o peso e a posição do jogador. Em outros casos, a escolha das informações não

é tão direta. Por exemplo, em um estudo da relação entre o estilo de aprendizagem preferido e a inteligência, como você definiria o estilo de aprendizagem e o mediria? Que medida de inteligência você usaria? É importante definir cuidadosamente as variáveis a serem estudadas e desenvolver métodos apropriados para determinar seus valores;

- Coleção de dados, que é a etapa de coleta de dados crucial, onde o pesquisador deve primeiro decidir se uma fonte de dados existente é adequada ou se novos dados devem ser coletados. Se for tomada a decisão de usar os dados existentes, é importante entender como os dados foram coletados e para que propósito, de forma que quaisquer limitações resultantes também sejam totalmente compreendidas. Se novos dados devem ser coletados, um plano cuidadoso deve ser desenvolvido, porque o tipo de análise que é apropriado e as conclusões subsequentes que podem ser tiradas dependem de como os dados são coletados;
- Sumarização de dados e análise preliminar. Depois que os dados são coletados, a próxima etapa geralmente é uma análise preliminar que inclui resumir os dados gráfica e numericamente, esta análise inicial fornece uma visão sobre características importantes dos dados e pode fornecer orientação na seleção de métodos apropriados para análises futuras;
- Análise formal de dados é uma etapa de análise de dados que requer que o pesquisador selecione e aplique métodos estatísticos. Muito deste livro é dedicado a métodos que podem ser usados para realizar esta etapa;
- Interpretação de resultados. Várias questões devem ser abordadas nesta etapa final. Alguns exemplos são: O que podemos aprender com os dados? Que conclusões podem ser tiradas da análise? Como nossos resultados podem guiar pesquisas futuras? A etapa de interpretação frequentemente leva à formulação de novas questões de pesquisa e essas novas questões levam de volta à primeira etapa. Dessa forma, uma boa análise de dados costuma ser um processo iterativo.

A Análise Exploratória de Dados pode não ocorrer de forma sequencial, sendo normal que a saída de uma etapa, faça com que seja necessário que outra etapa seja refeita e assim, seja garantida a qualidade da análise.

Outro ponto importante, é que a saída desta etapa de análise de dados será um *dataset* (conjunto de dados) pronto para servir de entrada para um modelo, que aí sim, irá aprender extraindo os padrões lá existentes.

2.3.1 Limpeza de dados e *Feature Engineering*

A *Feature Engineering* ou engenharia de recursos é o processo que cria novos recursos de entrada para o aprendizado de máquina, sendo uma das maneiras mais eficazes de melhorar os modelos preditivos. Através da *Feature engineering* / engenharia de recursos, você pode isolar as principais informações, destacar padrões e trazer conhecimentos de domínio (BRANTS *et al.*, 2007).

Brants *et al*. (2007), ainda complementam, mas o que é engenharia de recursos? A engenharia de recursos nada mais é do que um tópico informal onde há muitas definições possíveis. O fluxo de trabalho de aprendizado de máquina é fluido e iterativo; portanto, não há uma "resposta certa". Em poucas palavras, definimos engenharia de recursos como a criação de novos recursos a partir dos existentes para melhorar o desempenho do modelo.

Para Brants *et al.*,(2007), um processo típico de ciência de dados pode ser assim:

- Escopo do Projeto / Coleta de Dados*;*
- Análise Exploratória*;*
- Limpeza de Dados*;*
- Engenharia de Recursos*;*
- Treinamento do modelo (incluindo validação cruzada para ajustar os hiperparâmetros)*;*
- Entrega / Insights do Projeto.

A *Feature Engineering* desempenha um papel importante em muitas abordagens de classificação de textos clínicos, que envolve a seleção de um subconjunto de recursos informativos e a combinação de recursos distintos em novos recursos para obter uma representação que permita a classificação. No domínio de classificação de texto, os recursos normalmente incluem todos os termos distintos, como palavras, conceitos, presentes em um corpus de texto.

Brants *et al.* (2007) caracterizam que, mesmo corpos pequenos podem possuir dezenas de milhares de recursos, potencialmente necessitando de Feature engineering para

uma determinada tarefa de classificação. O conhecimento do domínio é frequentemente usado para orientar o processo de engenharia de recursos, como por exemplo, para identificar notas que afirmam a presença de uma doença onde os especialistas definem manualmente dicionários de termos relacionados à doença, como sintomas e medicamentos.

No entanto, a engenharia manual de recursos pode exigir um esforço considerável, e os recursos e grupos de recursos selecionados são problemas e específicos de domínio (GARLA; BRANDT, 2012).

Garla e Brandt (2012), ainda destacam que antes de dar início ao treinamento de um modelo, é preciso realizar a etapa de limpeza de dados, e muitas vezes *feature engineering* (engenharia de atributos). Isso se faz necessário em razão de modelos de *machine learning* ser uma representação dos padrões existentes nos dados, logo, eles precisam estar livres de imperfeições decorrentes do formato e/ou processo de coleta. Mas afinal o que é a Limpeza de dados? A Limpeza de dados consiste em tratar algumas imperfeições existentes nos dados, por exemplo, valores ausentes, valores fora de escala ou inválidos (ex: idade paciente = 200 anos), *outliers* ou valores extremos, mas que podem estar corretos, dados duplicados, transformar variáveis categóricas ou textuais em binárias, entre outros. Este processo é regrado em técnicas estatísticas que serão mencionadas mais a frente.

Para McKinney (2018)**,** uma parte importante do processo de desenvolvimento do modelo é chamada de engenharia de recursos no aprendizado de máquina, onde pode descrever qualquer transformação ou análise de dados que extraia informações de um conjunto de dados bruto que pode ser útil em um contexto de modelagem.

Complementando McKinney (2018**)** destaca que**,** um exemplo de *feature engineering* seria criar outro atributo com base em outros atributos existentes nos dados. Por exemplo, com base no total de consultas de um paciente em um determinado período de anos, criar um atributo que representa a média anual de consultas, este atributo provavelmente será mais relevante para um modelo de machine learning do que somente o total absoluto de consultas do paciente ao longo do tempo. Na visão de Domingos (2012)**,** um dos principais fatores de sucesso nos projetos de *machine learning* é a *feature engineering*.

Além disso, Domingos (2012) relata que a engenharia de recursos é a chave: no final do dia, alguns projetos de aprendizado de máquina são bem-sucedidos e outros falham. E a diferença se dá pelo fator mais importante: recursos usados. Domingos afirma ainda que, o aprendizado é fácil se você tiver muitos recursos independentes que se correlacionam bem com a classe. Por outro lado, se a classe for uma função muito complexa dos recursos, você

pode não ser capaz de aprendê-la. Frequentemente, os dados brutos não estão em um formato de fácil aprendizagem, mas você pode construir recursos a partir deles. Normalmente, é aqui que vai a maior parte do esforço em um projeto de aprendizado de máquina, e muitas vezes é também uma das partes mais interessantes, onde a intuição, criatividade e "arte negra" são tão importantes quanto às coisas técnicas.

2.4 ESTATÍSTICA DESCRITIVA E ESTATÍSTICA INFERENCIAL

Os métodos para organizar e resumir os dados, como o uso de tabelas, gráficos ou resumos numéricos, constituem o ramo da estatística denominado estatística descritiva.

Segundo McKinney (2018)**,** se tratando de estatísticas, descritiva e inferencial, quando apresentado pela primeira vez com um conjunto de medidas, seja uma amostra ou uma população, você precisa encontrar uma maneira de organizá-lo e resumi-lo. O ramo da estatística que apresenta técnicas para descrever conjuntos de medidas é chamado de estatística descritiva, que se apresenta em muitas formas: gráficos de barras, gráficos de setores e gráficos de linhas apresentados por um candidato político; tabelas numéricas no jornal; ou a quantidade média de precipitação relatada pelo meteorologista da televisão local. Gráficos gerados por computador e resumos numéricos são comuns em nossa comunicação diária.

Neste sentido McKinney (2018) enfatiza que a estatística descritiva se define por consistir em procedimentos usados para resumir e descrever as características importantes de um conjunto de medições.

Da mesma forma, como menciona McKinney (2018)**,** se o conjunto de medidas for a população inteira, você só precisa tirar conclusões com base nas estatísticas descritivas. No entanto, pode ser muito caro ou demorado enumerar toda a população. Talvez enumerar a população a destrua, como no caso dos testes do “tempo até a falha”. Por essas ou outras razões, você pode ter apenas uma amostra da população. Observando a amostra, você deseja responder a perguntas sobre a população como um todo. O ramo da estatística que lida com esse problema é chamada de estatística inferencial. Já a estatística inferencial se define por consistir em procedimentos usados que fazem inferências sobre as características da população a partir de informações contidas em uma amostra retirada dessa população. O objetivo é fazer inferências, ou seja, tirar conclusões, fazer previsões e tomar decisões sobre as características de uma população a partir das informações contidas em uma amostra.

Mendenhall, Beaver e Beaver (2019) descrevem como algumas medidas podem ajudar a obter uma melhor compreensão dos dados, um conjunto de dados com medidas numéricas onde os gráficos podem ajudá-lo a descrever a forma básica de uma distribuição de dados; "uma imagem vale mais que mil palavras." Existem limitações, no entanto, para o uso de gráficos. Suponhamos que você precise exibir seus dados para um grupo de pessoas e a lâmpada do projetor de dados apagar! Ou talvez você precise descrever seus dados por telefone, sem a possibilidade de exibir os gráficos! Você precisa encontrar outra maneira de transmitir uma imagem mental dos dados ao seu público. Uma segunda limitação é que os gráficos são um tanto imprecisos para uso em inferência estatística. Por exemplo, digamos que você queira usar um histograma de amostra para fazer inferências sobre um histograma de população. Como você pode medir as semelhanças e diferenças entre os dois histogramas de alguma forma concreta? Se eles fossem idênticos, você poderia dizer "Eles são iguais!" Mas, se eles forem diferentes, é difícil descrever o "grau de diferença". Uma maneira de superar esses problemas é usar medidas numéricas, que podem ser calculadas para uma amostra ou uma população de medidas. Você pode usar os dados para calcular um conjunto de números que transmitirá uma boa imagem mental da distribuição de frequência. Tais medidas são chamadas de parâmetros quando associadas à população e são chamadas de estatísticas quando calculadas a partir de medições de amostra. A definição das medidas descritivas numéricas associadas a uma população de medidas é chamada de parâmetros; aqueles calculados a partir de medições de amostra são chamados de estatísticas. Tem também a média aritmética de um conjunto de medidas que é uma medida de centro muito comum e útil. Esta medida é frequentemente referida como a média aritmética, ou simplesmente a média, de um conjunto de medidas. Para distinguir entre a média da amostra e a média da população, usaremos o símbolo x (x-bar) para uma média da amostra e o símbolo m (mu minúsculo do grego) para a média de uma população. Define-se que a média aritmética ou média de um conjunto de n medidas é igual a soma das medidas dividida por n ou que a mediana m de um conjunto de n medições é o valor de x que cai na posição intermediária quando as medições são ordenadas da menor para a maior. Estas medidas são fundamentais para entender como os dados estão distribuídos. Além disso, também podem ser utilizadas na etapa de limpeza de dados e engenharia de atributos, para imputar features ausentes.

2.4.1 Correlação - Medidas de Dispersão

Compreender como os dados variam em relação ao ponto central, é importante para entender relações entre variáveis e resultados obtidos. A melhor maneira de entender um desvio padrão é considerar o que as duas palavras significam (URDAN, 2015).

Ainda segundo Urdan (2015) o desvio, neste caso, refere-se à diferença entre uma pontuação individual em uma distribuição e a pontuação média da distribuição. Portanto, se a pontuação média de uma distribuição for 10 e uma criança individual tiver uma pontuação 12, o desvio será dois. A outra palavra no termo desvio padrão é padrão. Nesse caso, padrão significa típico ou médio. Portanto, um desvio padrão é o desvio típico, ou médio, entre as pontuações individuais em uma distribuição e a média da distribuição. Esta é uma estatística muito útil porque fornece uma medida útil de quão espalhadas as pontuações estão na distribuição. Quando combinados, a média e o desvio padrão fornecem uma imagem muito boa de como é a distribuição das pontuações. Desta forma a faixa fornece uma medida da propagação total em uma distribuição, nesse caso, da pontuação mais baixa para a mais alta, enquanto a variância e o desvio padrão são medidas da quantidade média de propagação dentro da distribuição. Os pesquisadores tendem a olhar para o intervalo quando desejam um instantâneo rápido de uma distribuição, como quando desejam saber se todas as categorias de resposta em uma pergunta da pesquisa foram usadas ou querem uma noção do equilíbrio geral das pontuações na distribuição. Os pesquisadores raramente olham apenas para a variância, porque ela não usa a mesma escala que a medida original de uma variável, embora a estatística de variância seja muito útil para o cálculo de outras estatísticas. O desvio padrão é uma estatística muito útil que os pesquisadores examinam constantemente para fornecer a medida mais facilmente interpretável e significativa da dispersão média das pontuações em uma distribuição.

Representado pela letra S: O desvio padrão ou desvio padrão populacional é uma medida de dispersão em torno da média populacional de uma variável aleatória. O termo possui também uma acepção específica no campo da estatística, na qual também é chamado de desvio padrão amostral representado pela letra **s** e indica uma medida de dispersão dos dados em torno de média amostral.

Resumo do desvio padrão: O desvio médio entre as pontuações individuais na distribuição e a média da distribuição.

Ainda conforme Urdan (2015), a distribuição normal é um conceito com o qual a maioria das pessoas tem alguma familiaridade, embora muitas vezes nunca tenham ouvido falar do termo, um nome mais familiar para a distribuição normal é a curva do sino, porque uma distribuição normal tem a forma de um sino. A distribuição normal é extremamente importante para as estatísticas e possui algumas características específicas que a tornam tão útil.

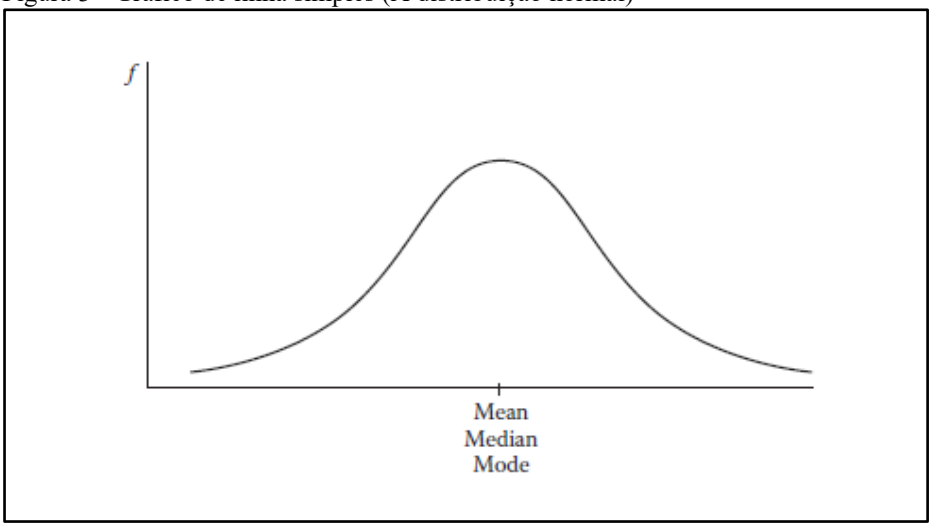
A Figura 3 apresenta um gráfico de linha simples que representa uma distribuição normal, esse tipo de gráfico mostra a frequência, o número de casos, com pontuações específicas em uma única variável. Portanto, neste gráfico, o eixo y mostra a frequência dos casos e o eixo x mostra a pontuação na variável de interesse. Por exemplo, se a variável fosse pontuações em um teste de QI, o eixo x teria as pontuações variando do menor ao maior. A média, mediana e moda seriam 100, e o pico da linha mostra que a frequência de casos é mais alta em 100. Conforme você se afasta do modo em qualquer direção, a altura da linha diminui, indicando menos casos (frequências mais baixas) nessas outras pontuações.

Da mesma forma a Figura 3, mostra ainda que a distribuição normal tem três características fundamentais. Em primeiro lugar, é simétrico, o que significa que a metade superior e a metade inferior da distribuição são imagens espelhadas uma da outra. Em segundo, apresenta que a média, a mediana e a moda estão todos no mesmo lugar, no centro da distribuição (no topo da curva do sino). Por causa dessa segunda característica, a distribuição normal é mais alta no meio, é unimodal e se curva para baixo em direção ao topo e à base da distribuição. Finalmente, a distribuição normal é assintótica, o que significa que as caudas superior e inferior da distribuição nunca realmente tocam a linha de base, também conhecida como eixo x.

Mas por que a distribuição normal é tão importante? Pois bem, quando os pesquisadores coletam dados de uma amostra, tudo o que querem saber são as características da amostra. Por exemplo, se quiséssemos examinar os hábitos alimentares de 100 estudantes universitários do primeiro ano, apenas selecionaria 100 alunos, perguntaria o que eles comem e resumiria os dados. Tais dados podem me fornecer estatísticas como o número médio de calorias consumidas por dia pelos 100 alunos da minha amostra, os alimentos mais consumidos, a variedade de alimentos consumidos e assim por diante. Todas essas estatísticas simplesmente descrevem as características da minha amostra e, portanto, são chamadas de estatísticas descritivas, que geralmente são usadas apenas para descrever uma amostra

específica. Quando tudo o que nos interessa é descrever uma amostra específica, não importa se os escores da amostra são normalmente distribuídos ou não (URDAN, 2015).

Figura 3 - Gráfico de linha simples (A distribuição normal)



Finte: URDAN (2015, p. 30).

Porém se há necessidade de fazer mais do que simplesmente descrever uma amostra, deve-se saber qual é a probabilidade exata de algo ocorrer em sua amostra apenas por acaso. Por exemplo, se o aluno médio em minha amostra consome 2.000 calorias por dia, quais são as chances, ou probabilidade, de ter um aluno na amostra que consome 5.000 calorias por dia? Cada uma das três características da distribuição normal é crítica em estatística porque nos permite fazer bom uso da estatística de probabilidade. Além disso, é possível fazer inferências sobre a população com base nos dados que coletam de sua amostra. Para determinar se algum fenômeno observado em uma amostra representa um fenômeno real na população da qual a amostra foi extraída, estatísticas inferenciais são usadas. Por exemplo, suponhamos que na população de homens e mulheres não haja diferença no número médio de calorias consumidas por dia. Tal suposição de que não há diferenças é conhecida como hipótese nula. Agora, vamos supor que eu selecione uma amostra de homens e uma amostra de mulheres, compare seu consumo calórico diário médio e descubra que os homens comem em média 200 calorias a mais por dia do que as mulheres. Dada hipótese nula de nenhuma diferença, qual é a probabilidade de encontrar por acaso uma diferença tão grande entre as amostras? Para calcular essas probabilidades, é necessário confiar na distribuição normal, porque as características da distribuição normal permitem que os estatísticos gerem estatísticas de probabilidade exatas (MARASCUILO; SERLIN, 1988).

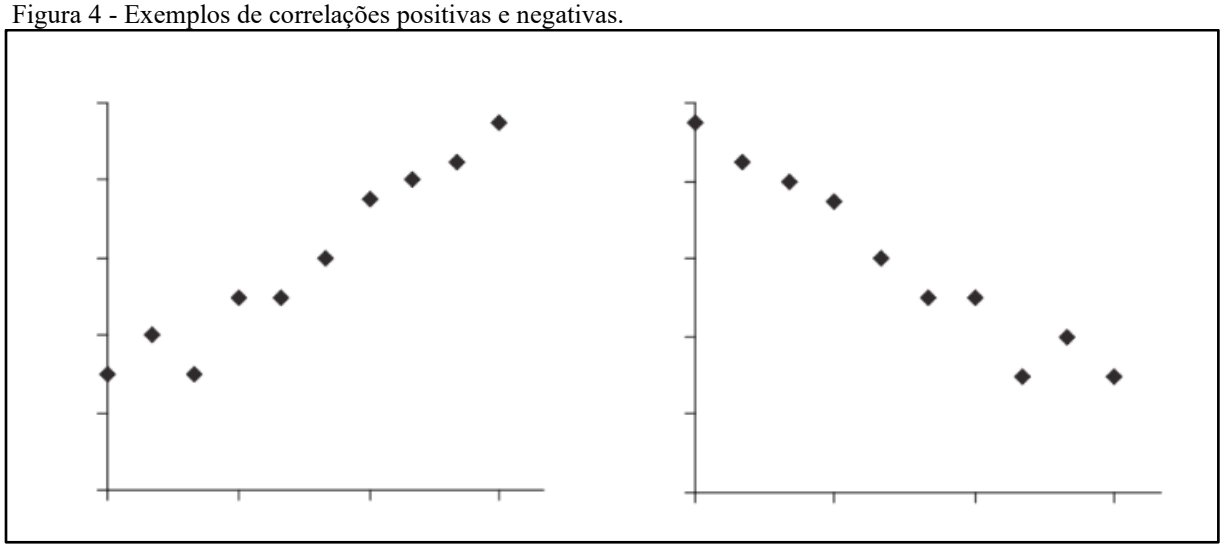
A partir de agora o tema correlação será abordado, embora haja vários tipos diferentes de coeficientes, o mais comumente usado na pesquisa em ciências sociais é o coeficiente de correlação momento-produto de *Pearson*. Onde os pesquisadores calculam os coeficientes de correlação quando desejam saber como duas variáveis estão relacionadas entre si. Para uma correlação produto-momento de *Pearson*, ambas as variáveis devem ser medidas em uma escala de intervalo ou razão e são conhecidas como variáveis contínuas. Por exemplo, suponhamos que queremos saber se há uma relação entre a quantidade de tempo que os alunos passam estudando para um exame e suas notas no exame. Suspeitando ainda, que quanto mais horas os alunos gastam estudando, maior será sua pontuação no exame. Mas também, muito provável que não haja uma correspondência perfeita entre o tempo gasto estudando e os resultados dos testes, e alguns alunos obterão notas baixas no exame, mesmo que estudem por muito tempo, simplesmente porque podem ter dificuldade em entender o material. Na verdade, haverá vários alunos que passarão um período excessivamente longo de tempo estudando para o teste, precisamente porque estão tendo problemas para entender o material. Por outro lado, provavelmente haverá alguns alunos que se saem muito bem no teste sem gastar muito tempo estudando. Apesar dessas "exceções" à regra, ainda supondo que, em média, à medida que aumenta a quantidade de tempo gasto estudando, o mesmo acontece com as notas dos alunos no exame (COHEN *et al.*, 2013).

Conforme Urdan (2015) existem duas características fundamentais dos coeficientes de correlação com os quais os pesquisadores se preocupam, o primeiro deles é a direção do coeficiente de correlação, nos quais podem ser positivos ou negativos. Uma correlação positiva indica que os valores nas duas variáveis que estão sendo analisadas se movem na mesma direção, ou seja, conforme as pontuações em uma variável aumentam as pontuações na outra variável também aumentam (em média). Da mesma forma, em média, à medida que as pontuações em uma variável diminuem, as pontuações na outra variável diminuem. Portanto, se houver uma correlação positiva entre a quantidade de tempo que os alunos passam estudando e suas pontuações no teste, pode-se que, em média, quanto mais tempo os alunos passam estudando, mais altas são suas pontuações no teste. Isto é equivalente a dizer que, o menor tempo que passam estudando, mais baixa serão suas pontuações no teste. Ambos representam uma positiva correlação entre o tempo gasto estudando e os resultados dos testes.

Já uma correlação negativa indica que os valores das duas variáveis em análise avançam de maneiras opostas, ou seja, conforme a pontuação de uma variável aumenta, a pontuação da outra variável diminui e vice-versa (em média). Se houvesse uma correlação

negativa entre a quantidade de tempo gasto estudando e as notas dos testes, saberíamos que, em média, quanto mais tempo os alunos passam estudando para o exame, menor seria a pontuação deles no exame. Da mesma forma, com uma correlação negativa, pode-se concluir que, o menor tempo que os alunos passam estudando, o superior suas pontuações estão no exame. Essas correlações positivas e negativas são representadas por diagramas de dispersão conforme se pode visualizar na Figura 4. Os diagramas de dispersão são simplesmente gráficos que indicam as pontuações de cada caso em uma amostra simultaneamente em duas variáveis. Por exemplo, no diagrama de dispersão de correlação positiva da Figura 4, o primeiro caso da amostra estudou por 1 hora e obteve pontuação 30 no exame. O segundo caso estudou por duas horas e obteve uma nota 40 no exame.

Figura 4 - Exemplos de correlações positivas e negativas.



Fonte: URDAN (2015, p.80).

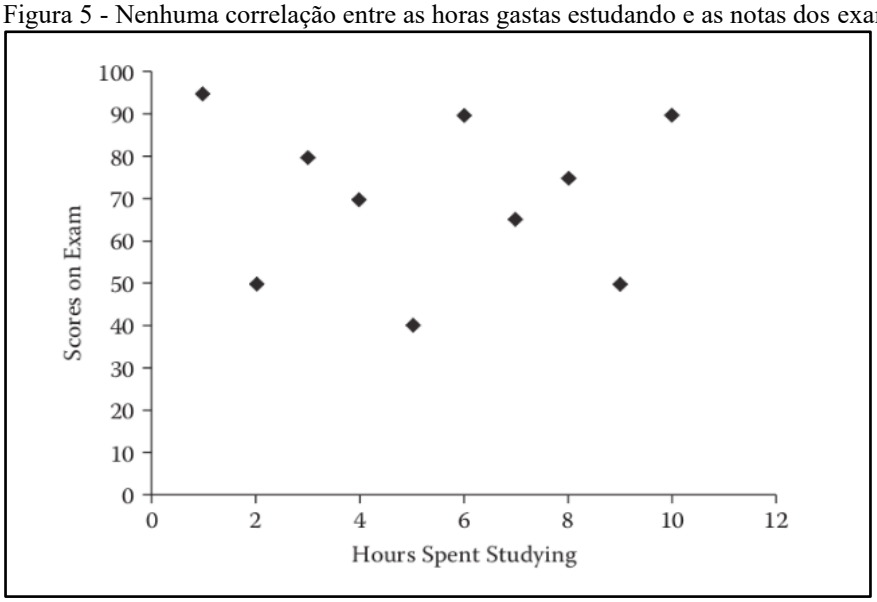
Na visão de Urdan (2015) uma característica fundamental dos coeficientes de correlação é a força ou grandeza e magnitude do relacionamento, nos quais os coeficientes de correlação variam em intensidade de -1,00 a +1,00. Um coeficiente de correlação de .00 indica que não há relação entre as duas variáveis que estão sendo examinadas, ou seja, os escores de uma das variáveis não estão relacionados de forma significativa aos escores da segunda variável. Quanto mais próximo o coeficiente de correlação estiver de -1,00 ou +1,00, mais forte será a relação entre as duas variáveis. Uma correlação negativa perfeita de -1,00 indica que para cada membro da amostra ou população, uma pontuação mais alta em uma variável está relacionada a uma pontuação mais baixa na outra variável. Já uma correlação positiva perfeita de +1,00 revela que para cada membro da amostra ou população, uma

pontuação mais alta em uma variável está relacionada a uma pontuação mais alta na outra variável.

Como descrito por Urdan (2015), correlações perfeitas nunca são encontradas na pesquisa real das ciências sociais. Geralmente, os coeficientes de correlação ficam entre –.70 e +.70. Alguns autores sugerem que os coeficientes de correlação entre – .20 e +.20 indicam uma relação fraca entre duas variáveis, aquelas entre .20 e 0,50 (positivo ou negativo) representam uma relação moderada, e aqueles maiores que .50 (positivo ou negativo) representam um relacionamento forte. Essas regras gerais para julgar a relevância dos coeficientes de correlação devem ser consideradas com cautela. Por exemplo, mesmo uma "pequena" correlação entre o consumo de álcool e doença hepática (por exemplo, +.15) é importante, enquanto uma forte correlação entre o quanto as crianças gostam de sorvete de baunilha e chocolate (por exemplo, +.70) pode não seja tão importante.

Urdan (2015) ainda complementa que, os diagramas de dispersão apresentados, na Figura 4, representam correlações positivas e negativas muito fortes (r =. 97 e r = –,97 para as correlações positivas e negativas, respectivamente; r é o símbolo do coeficiente de correlação de Pearson da amostra). Na Figura 5, é apresentado um diagrama de dispersão que representa virtualmente nenhuma correlação entre o número de horas gastas estudando e as pontuações no exame. Observe que não há um padrão discernível entre as pontuações nas duas variáveis. Em outras palavras, os dados apresentados na Figura 5 revelam que seria virtualmente impossível prever a pontuação do teste de um indivíduo simplesmente sabendo quantas horas a pessoa estudou para o exame.

Figura 5 - Nenhuma correlação entre as horas gastas estudando e as notas dos exames.



Fonte: URDAN (2015).

Para Urdan (2015) existem várias fórmulas diferentes que podem ser usadas para calcular os coeficientes de correlação produto-momento de Pearson. Essas fórmulas produzem o mesmo resultado e diferem apenas na facilidade de uso. Na verdade, nenhum deles é particularmente fácil de usar.

A fórmula apresentada na Figura 6 requer padronização de suas variáveis, onde se está simplesmente subtraindo a média de cada pontuação em sua amostra e dividindo pelo desvio padrão, fornecendo assim um *z Ponto* para cada caso na amostra. Os membros da amostra com pontuações abaixo da média terão resultados negativos z pontuações, enquanto os membros da amostra com pontuações acima da média terão z pontuações.

Por exemplo, o denominador é N, que é o número de pares de pontuações (ou seja, o número de casos na amostra). Sempre que dividimos por N, estamos encontrando uma média. Portanto, sabemos que o coeficiente de correlação será uma média de algum tipo. Ao olhar o numerador vemos que devemos encontrar a soma (Σ) de alguma coisa. Mas na fórmula para calcular o coeficiente de correlação, temos que encontrar a soma dos produtos cruzados Entre o *z pontuações* em cada uma das duas variáveis que estão sendo examinadas para cada caso na amostra. Quando multiplicada a pontuação de cada indivíduo em uma variável com a pontuação desse indivíduo na segunda variável (ou seja, encontrar um produto vetorial), some aquele em todos os indivíduos, em seguida, divida por N, temos um produto cruzado médio, conhecido como covariância. Se padronizar essa covariância, acabaremos com um coeficiente de correlação. Na fórmula fornecida na Figura 6, aparece à padronização das variáveis antes de calcular os produtos cruzados, produzindo assim uma estatística de covariância, que é um coeficiente de correlação.

Se um caso individual na amostra tiver pontuações acima da média em cada uma das duas variáveis sendo examinadas, as pontuações dos dois z sendo multiplicadas serão positivas e o produto vetorial resultante também será positivo. Da mesma forma, se um caso individual tiver pontuações abaixo da média em cada uma das duas variáveis, o z das pontuações multiplicadas será negativo e o produto vetorial será novamente positivo. Portanto, se tivermos uma amostra em que pontuações baixas em uma variável tendem a estar associadas a pontuações baixas na outra variável e pontuações altas em uma variável tendem a estar associadas a pontuações altas na segunda variável, então, quando somamos os produtos de nossas multiplicações, vamos acabar com um número positivo. É assim que se obtém um coeficiente de correlação positivo. E se um caso individual em uma amostra tem uma pontuação que é maior do que a média na primeira variável (ou seja, um resultado positivo z

pontuação) e uma pontuação que está abaixo da média na segunda variável (ou seja, uma pontuação negativa z pontuação), quando estes dois z pontuações são multiplicadas juntas, elas irão produzir um negativo produto. Se, para a maioria dos casos da amostra, pontuações altas em uma variável estão associadas a pontuações baixas na segunda variável, a soma dos produtos do z pontuações [Σ (z x z y)] será negativo. Então assim obtemos um coeficiente de correlação negativo.

Figura 6 - Tabela Fórmula de definição para coeficiente de correlação de Pearson

$$r = \frac{\Sigma (z_x z_y)}{N}$$

Onde

r = Coeficiente de correlação produto-momento de Pearson

z_x = uma z pontuação para variável X

z_y = um par z pontuação para variável Y N = o número de pares de X e Y pontuações

Fonte: Adaptado de URDAN (2015, p. 81).

- Correlações perfeitas, ou seja, entre 1 e -1, geralmente não ocorrem;
- Até 0,2 ou < -0,2 fraca
- > 0,2 e <= 0,5 < ou > -0,2 e <= -0,5 moderada
- > 0,5 ou < -0,5 forte

2.4.2 Correlação não é causa

Coeficientes de correlação, como o de *Pearson,* são estatísticas muito poderosas. Eles nos permitem determinar se, em média, os valores de uma variável estão associados aos valores de uma segunda variável. Essa pode ser uma informação muito útil, mas as pessoas, incluindo cientistas sociais, muitas vezes são tentadas a atribuir mais significado aos coeficientes de correlação do que merecem. Ou seja, as pessoas costumam confundir os conceitos de correlação e causalidade. Correlação (correlação) significa simplesmente que a variação nas pontuações de uma variável corresponde à variação nas pontuações de uma segunda variável. Causação significa que a variação nas pontuações de uma variável causa ou cria variação nas pontuações de uma segunda variável (PEDHAZUR, 1982)

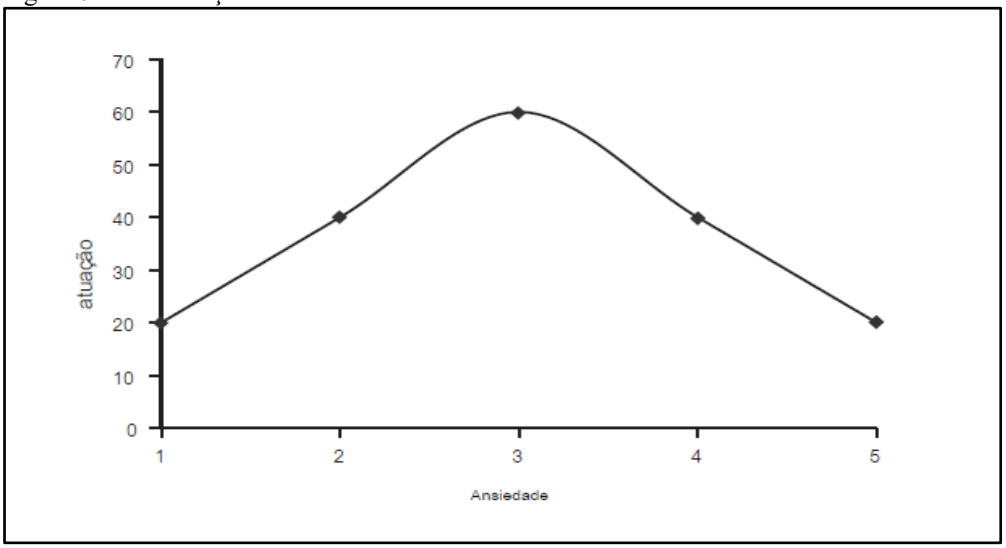
De acordo com Urdan (2015) O coeficiente de correlação, como o de *Pearson*, são estatísticas muito poderosas, que nos permitem determinar se os valores de uma variável são

associados com os valores em uma segunda variável. Essa pode ser uma informação muito útil, mas as pessoas, incluindo cientistas sociais, muitas vezes são tentadas a atribuir mais significado aos coeficientes de correlação do que merecem, sendo assim, as pessoas costumam confundir os conceitos de correlação e causalidade.

Correlação (co-relação) significa simplesmente que a variação nas pontuações em uma variável corresponde com variação nas pontuações em uma segunda variável. Já a Causação significa que a variação nas pontuações em uma variável causa ou cria variação nas pontuações em uma segunda variável. Além da questão da correlação-causalidade existe algumas outras características importantes das correlações que vale a pena observarem. Primeiramente, as correlações simples de *Pearson* são projetadas para examinar as relações lineares entre as variáveis, em outras palavras, eles descrevem relações retas médias entre variáveis. Por exemplo, se você encontrar uma correlação positiva entre duas variáveis poderá prever quanto às pontuações em uma variável aumentarão com cada aumento correspondente na segunda variável, mas nem todas as relações entre variáveis são lineares. Por exemplo, existe uma relação curvilínea entre ansiedade e desempenho em vários comportamentos acadêmicos e não acadêmicos, ao fazer um teste de matemática, um pouco de ansiedade pode realmente ajudar no desempenho, no entanto, quando o aluno fica muito nervoso, essa ansiedade pode interferir no desempenho.

Chamamos isso de relação curvilínea porque o que começou como uma relação positiva entre desempenho e ansiedade em níveis mais baixos de ansiedade torna-se uma relação negativa em níveis mais altos de ansiedade. Essa relação curvilínea é apresentada graficamente na Figura 7. Como os coeficientes de correlação mostram a relação média entre duas variáveis, quando a relação entre duas variáveis é curvilínea, o coeficiente de correlação pode ser bem pequeno, sugerindo uma relação mais fraca do que pode realmente existir (JACCARD; TURRISI; WAN, 1990).

Figura 7 - Uma relação curvilínea.



Fonte: URDAN (2015, p.84).

2.5 ALGORITMOS DE ÁRVORE, CLASSIFICAÇÃO E REGRESSÃO

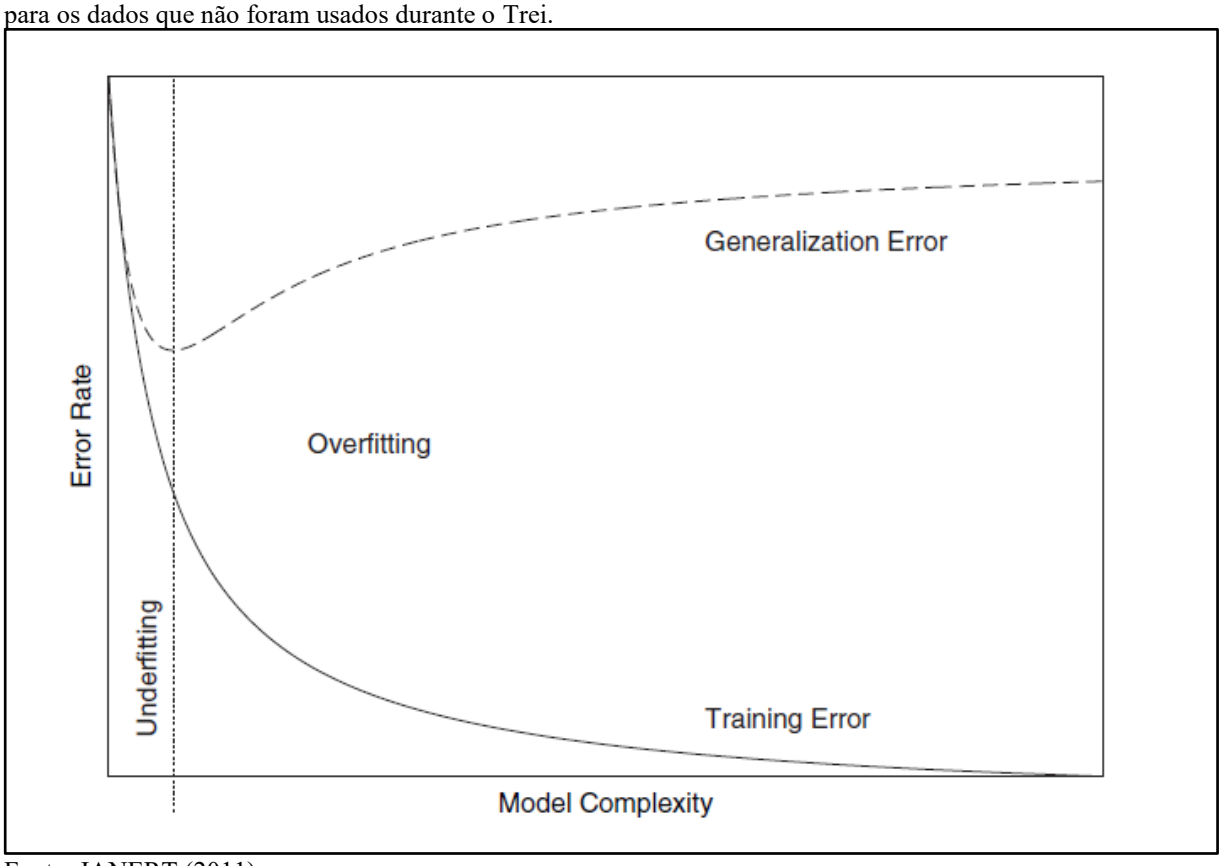
Uma máquina aprende quando é capaz de acumular experiência por meio de dados, programas e desenvolver novos conhecimentos para que seu desempenho em tarefas específicas melhore com o tempo. Esta ideia de aprender com a experiência é central para os vários tipos de problemas encontrados no aprendizado de máquina, especialmente problemas envolvendo classificação como, por exemplo, reconhecimento de dígitos manuscritos, reconhecimento de fala, reconhecimento de rosto, classificação de texto. O objetivo geral de cada um desses problemas é encontrar uma maneira sistemática de classificar um exemplo futuro (por exemplo, uma amostra de caligrafia, uma palavra falada, uma imagem de rosto, um fragmento de texto). A classificação é baseada em medições naquele exemplo futuro junto com o conhecimento obtido de uma amostra de aprendizagem (ou treinamento) de exemplos semelhantes (onde a classe de cada exemplo é completamente determinada e conhecida, e o número de classes é finito e conhecido) (VAPNIK, 2000).

Vapnik (2000) enfatiza que a necessidade de criar novos métodos e terminologia para analisar grandes e complexos conjuntos de dados levou pesquisadores de várias disciplinas de estatística, reconhecimento de padrões, redes neurais, aprendizagem de máquina simbólica, teoria de aprendizagem computacional e, é claro, IA para trabalharem juntos para influenciar o desenvolvimento de aprendizado de máquina.

Entre as técnicas usadas para resolver problemas de aprendizado de máquina, os tópicos de maior interesse para os estatísticos - estimativa de densidade, regressão e

reconhecimento de padrões (incluindo redes neurais, análise discriminante, classificadores baseados em árvore, florestas aleatórias, ensacamento e reforço, máquinas de vetor de suporte, clustering e métodos de redução de dimensionalidade) - agora são chamados coletivamente de aprendizado estatístico. O aprendizado de máquina divide os problemas de aprendizado em várias categorias, mas as duas mais relevantes para a estatística são o aprendizado supervisionado e o aprendizado não supervisionado. De maneira breve, a aprendizagem supervisionada recebe um conjunto de variáveis de entrada contínuas ou categóricas e uma variável de saída correta (que é observada ou fornecida por um "professor" explícito) e tenta encontrar uma função das variáveis de entrada para aproximar as conhecidas variáveis de saída: uma variável de saída contínua produz um problema de regressão, enquanto uma variável de saída categórica produz um problema de classificação. E na aprendizagem não supervisionada não há informações disponíveis (ou seja, nenhum "professor" explícito) para definir uma variável de saída apropriada; muitas vezes referido como "descoberta científica" (VAPNIK, 2000).

Figura 8 - *Overfitting*: conforme um modelo se torna mais complexo, ele se torna cada vez mais capaz de representar os dados de treinamento. No entanto, esse modelo é excessivamente ajustado e não generalizará bem para os dados que não foram usados durante o Trei.



Fonte: JANERT (2011).

Para entender a necessidade de uma fase de teste separada (usando um conjunto de dados separado), deve-se ter em mente que, desde que usemos parâmetros suficientes (ou seja, tornando o classificador cada vez mais complexo), podemos sempre ajustar um classificador até que ele funcione perfeitamente bem no conjunto de treinamento. Mas, ao fazer isso, treinamos o classificador para memorizar todos os aspectos do conjunto de treinamento, incluindo aqueles que são atípicos para o sistema em geral. Portanto, precisamos encontrar o nível certo de complexidade para o classificador. Por um lado, se for muito simples, não poderá representar muito bem o comportamento desejado e tanto seu erro de treinamento como de generalização serão pobres; isso é conhecido como *underfitting* (sob ajuste). Por outro lado, se tornarmos o classificador muito complexo, ele terá um desempenho muito bom no conjunto de treinamento (erro de treinamento baixo), mas não generalizará bem para pontos de dados desconhecidos (erro de generalização alto); isso é conhecido como *overfitting* (sobreajuste) (IZENMAN, 2008). A Figura 8 resume esses conceitos.

Depois que um classificador foi desenvolvido e testado, ele pode ser usado para classificar pontos de dados verdadeiramente novos e desconhecidos, ou seja, pontos de dados para os quais o rótulo de classe correto não é conhecido. Isso está em contraste com o conjunto de teste, onde os rótulos de classe eram conhecidos, mas não usados pelo classificador ao fazer uma previsão.

2.5.1 Algoritmos de Classificação

Foram desenvolvidas, ao menos, meia dúzia de famílias diferentes de algoritmos de classificação.

2.5.1.1. Classificadores baseados em instância e métodos vizinhos mais próximos

Os Classificadores baseados em instância têm por objetivo classificar uma instância desconhecida, e/ou simplesmente encontrar uma instância existente que seja “mais semelhante” à nova instância e atribuir o rótulo de classe da instância conhecida à nova. Essa ideia básica pode ser generalizada de várias maneiras. Em primeiro lugar, a noção de “mais semelhante” apresenta a noção de distância e medidas de similaridade, obviamente, temos uma flexibilidade considerável na escolha de qual medida de distância usar. Além disso, não temos que parar em uma única instância existente "mais semelhante". Em vez disso, podemos

pegar os k vizinhos mais próximos e usá-los para classificar a nova instância, normalmente usando uma regra da maioria (ou seja, atribuímos a nova instância à classe que ocorre com mais frequência entre os k vizinhos). Poderíamos ainda empregar a regra da maioria, na qual vizinhos “mais semelhantes” contribuem mais fortemente do que os mais distantes (IZENMAN, 2008).

Ainda segundo Izenman (2008), classificadores baseados em instância são atípicos porque não têm uma fase de treinamento, por esse motivo, eles também são conhecidos como “alunos preguiçosos”. (O único parâmetro ajustável é a extensão k da vizinhança usada para classificação). No entanto, um conjunto de instâncias conhecidas deve ser mantido disponível durante a fase final de aplicação, simplesmente pela classificação poder ser relativamente cara e porque o conjunto de instâncias existentes deve ser pesquisado em busca de vizinhos apropriados.

Izenman (2008), completa que, os classificadores baseados em instância são locais, eles não levam em consideração a distribuição geral dos pontos. Além disso, eles não impõem nenhuma forma ou geometria particular aos limites de decisão que geram. Nesse sentido, eles são especialmente flexíveis. Por outro lado, também são suscetíveis a ruídos. Por fim, os classificadores baseados em instância dependem da escolha adequada da medida de distância, assim como os algoritmos de agrupamento.

2.5.1.2. Classificadores Bayesianos

Um classificador *bayesiano* tem uma visão probabilística da classificação. Dado um conjunto de atributos, ele calcula a probabilidade da instância pertencer a esta ou aquela classe. Uma instância é então atribuída ao rótulo de classe com a maior probabilidade (JANERT, 2011).

Ainda segundo Janert (2011) um classificador bayesiano calcula uma probabilidade condicional, Figura 9, esta é a probabilidade de a instância pertencer a uma classe C específica, dado o conjunto de valores de atributos:

Figura 9 - Cálculo para probabilidade Condicional

$$P(\text{class } C \mid \{x_1, x_2, x_3, \ldots, x_n\})$$

Fonte: JANERT (2011, p. 409).

Aqui, C é o rótulo da classe e o conjunto de valores de atributos é {x1, x2, x3,. . . , xn}. Observe que ainda não sabemos o valor da probabilidade - se soubéssemos, estaríamos acabados (JANERT, 2011).

Para fazer progresso, segundo Janert (2011), invoca-se o teorema de *Bayes*, Figura 10, para inverter esta expressão de probabilidade da seguinte forma:

Figura 10 - Teorema de Bayes

$$P(\text{class } C \mid \{x_i\}) = \frac{P(\{x_i\} \mid \text{class } C) \cdot P(\text{class } C)}{P(\{x_i\})}$$

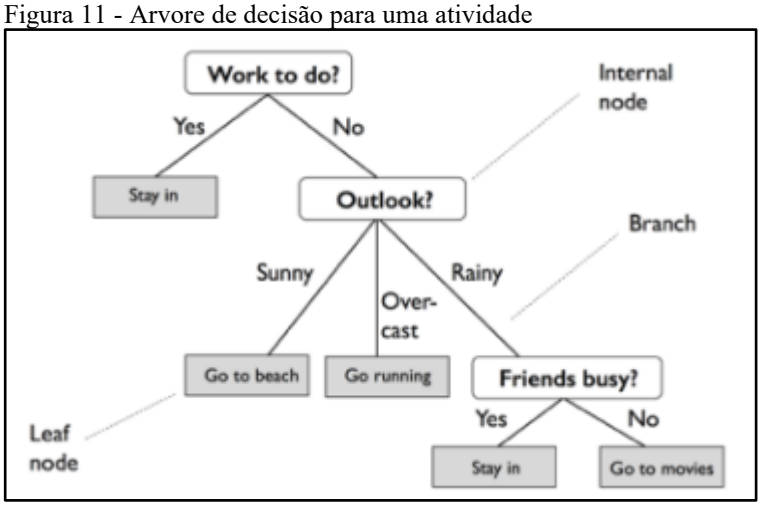
Fonte: JANERT (2011, p.410).

Onde foi reduzido o conjunto de n recursos em {xi} para abreviar.

2.5.1.3. Árvores de Decisão

Conforme Raschka e Mirjalili (2017) os classificadores de árvore de decisão são modelos atraentes. Como o nome árvore de decisão sugere, pode-se pensar neste modelo como dividindo nossos dados ao tomar uma decisão com base em uma série de perguntas Figura 11.

Figura 11 - Arvore de decisão para uma atividade



Fonte: RASCHKA; MIRJALILI (2017, p.89),

Com base nos recursos de nosso conjunto de treinamento, o modelo de árvore de decisão aprende uma série de perguntas para inferir os rótulos de classe das amostras. Embora

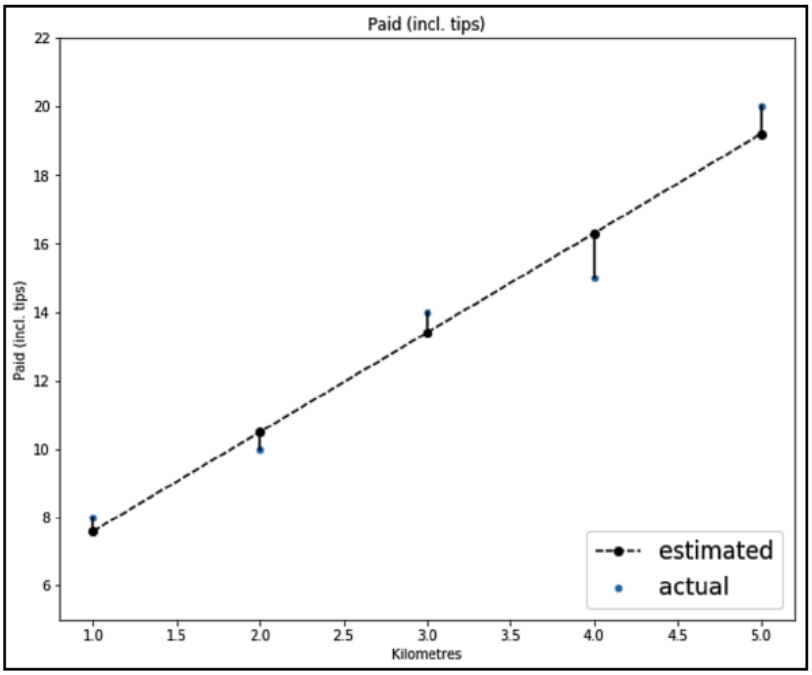
a figura anterior ilustre o conceito de uma árvore de decisão com base em variáveis categóricas, o mesmo conceito se aplica se nossos recursos forem números reais, como no conjunto de dados Iris. Exemplificando, poderíamos apenas definir um valor de corte ao longo do eixo da característica da largura da sépala e fazer uma pergunta binária "A largura da sépala é ≥ 2,8 cm?" (RASCHKA; MIRJALILI, 2017).

Usando o algoritmo de decisão, começamos na raiz da árvore e dividimos os dados no recurso que resulta no maior Ganho de Informação (IG). Em um processo iterativo, podemos repetir esse procedimento de divisão em cada nó até que as folhas estejam puras, sendo assim, todas as amostras em cada nó pertencem à mesma classe. Na prática, isso pode resultar em uma árvore muito profunda com muitos nós, o que pode facilmente levar ao sobreajuste. Normalmente queremos podar a árvore definindo um limite para a profundidade máxima da árvore (RASCHKA; MIRJALILI, 2017).

2.5.2 Algoritmos de regressão Linear

De acordo com Amr (2020) algoritmos têm tudo a ver com objetivos, e este objetivo não é viável se não existir uma relação linear entre os pontos. O algoritmo de regressão linear tenta encontrar uma linha onde a média dos erros quadráticos entre os pontos estimados na linha e os pontos reais sejam mínimos.

Figura 12 – Algoritmo Linear regressão



Fonte: AMR (2020, p. 66).

Conforme a Figura 12, o método usado para encontrar uma linha que minimiza o quadrado médio, o erro (MSE) é conhecido como mínimos quadrados ordinários. Frequentemente, a regressão linear significa apenas mínimos quadrados ordinários, se o método dos mínimos quadrados ordinários é usado ou um método diferente está sendo empregado. O método dos mínimos quadrados ordinários tem cerca de dois séculos e usa matemática simples para estimar os parâmetros, é por isso que alguns podem argumentar que esse algoritmo não é realmente um algoritmo de aprendizado de máquina. Amr (2020) segue uma abordagem mais liberal ao categorizar o que é aprendizado de máquina e o que não é. Contanto que o algoritmo aprende automaticamente com os dados e são usados esses dados para avaliá-los. Para o autor isso se enquadra no paradigma do aprendizado de máquina.

2.5.3 *Random Forest Classifier*

Para Raschka e Mirjalili (2017) *Random Forest* é uma variação de algoritmos de árvore. As florestas aleatórias ganharam enorme popularidade em aplicativos de aprendizado de máquina durante a última década devido ao seu bom desempenho de classificação, escalabilidade e facilidade de uso. Intuitivamente, uma floresta aleatória pode ser considerada um conjunto de árvores de decisão. A ideia por trás de uma floresta aleatória é fazer a média de árvores de decisão múltiplas profundas que individualmente sofrem de alta variância, para construir um modelo mais robusto que tem um melhor desempenho de generalização e é menos suscetível a *overfitting,* que nada mais é do que sobreajuste, um termo usado na estatística para descrever quando um modelo se ajusta bem ao conjunto de dados, porém ineficaz para prever resultados. O algoritmo de floresta aleatório pode ser resumido em quatro etapas simples:

- Desenhe uma amostra de *bootstrap* aleatória de tamanho n (escolha aleatoriamente n amostras do conjunto de treinamento com substituição);
- Aumente uma árvore de decisão a partir da amostra de *bootstrap*. Em cada nó:
- Selecione aleatoriamente d recursos sem substituição;
- Divida o nó usando o recurso que fornece a melhor divisão de acordo com a função objetivo, por exemplo, maximizar o ganho de informação;
- Repita as etapas 1-2 k vezes;

- Agregue a previsão por cada árvore para atribuir o rótulo da classe por maioria de votos.

Devemos observar uma pequena modificação na etapa 2 (dois) quando estamos treinando as árvores de decisão individuais: em vez de avaliar todos os recursos para determinar a melhor divisão em cada nó, consideramos apenas um subconjunto aleatório desses (RASCHKA; MIRJALILI, 2017).

Ainda conforme Raschka e Mirjalili (2017), caso você não esteja familiarizado com os termos amostragem com e sem substituição, vamos percorrer um experimento de pensamento simples. Vamos supor que estejamos jogando um jogo de loteria em que tiramos números aleatoriamente de uma urna. Começamos com uma urna que contém cinco números únicos, 0, 1, 2, 3 e 4, e desenhamos exatamente um número a cada volta. Na primeira rodada, a chance de tirar um determinado número da urna seria de 1/5. Agora, na amostragem sem reposição, não colocamos o número de volta na urna após cada volta. Consequentemente, a probabilidade de tirar um determinado número do conjunto de números restantes na próxima rodada depende da rodada anterior.

Raschka e Mirjalili (2017) destaca o exemplo em que, se tiver um conjunto restante de números 0, 1, 2 e 4, a chance de tirar o número 0 será 1/4 no próximo turno. No entanto, na amostragem aleatória com reposição, sempre devolvemos o número sorteado à urna, de modo que as probabilidades de tirar um determinado número a cada volta não mudem; podemos desenhar o mesmo número mais de uma vez. Em outras palavras, na amostragem com reposição, as amostras (números) são independentes e possuem covariância zero. Por exemplo, os resultados de cinco rodadas de sorteio de números aleatórios podem ser assim:.

- Amostragem aleatória sem substituição: 2, 1, 3, 4, 0;
- Amostragem aleatória com substituição: 1, 3, 3, 4, 1.

Embora as florestas aleatórias não ofereçam o mesmo nível de interpertabilidade que as árvores de decisão, uma grande vantagem das florestas aleatórias é que não precisamos nos preocupar tanto em escolher bons valores de hiperparâmetros. Normalmente, não precisamos podar a floresta aleatória, pois o modelo de conjunto é bastante robusto ao ruído das árvores de decisão individuais, o único parâmetro com o qual realmente precisamos nos preocupar na prática é o número de árvores k, que escolhemos para a floresta aleatória, onde, quanto maior

o número de árvores, melhor será o desempenho do classificador de floresta aleatório à custa de um custo computacional maior. Embora seja menos comum na prática, outros hiperparâmetros do classificador de floresta aleatório, que podem ser otimizados, compactando dados por meio da redução de dimensionalidade. Por meio do tamanho da amostra n da amostra de *bootstrap*, controlamos a compensação de viés-variância da floresta aleatória e então, diminuir o tamanho da amostra de bootstrap aumenta a diversidade entre as árvores individuais, uma vez que a probabilidade de que uma amostra de treinamento particular seja incluída na amostra de bootstrap é menor. Sendo assim, reduzir o tamanho das amostras de bootstrap pode aumentar a aleatoriedade da floresta aleatória e pode ajudar a reduzir o efeito do sobreajuste. No entanto, amostras de bootstrap menores normalmente resultam em um desempenho geral inferior da floresta aleatória, uma pequena lacuna entre o desempenho de treinamento e teste, mas um desempenho geral de teste baixo. Por outro lado, aumentar o tamanho da amostra de bootstrap pode aumentar o grau de sobreajuste. Como os exemplos de *bootstrap* e, consequentemente, as árvores de decisão individuais se tornam mais semelhantes entre si, eles aprendem a se ajustar ao conjunto de dados de treinamento original mais de perto (RASCHKA; MIRJALILI, 2017).

Neste sentido Raschka e Mirjalili (2017) complementam que, na maioria das implementações, incluindo a implementação de Random Forest Classifier no scikit-learn, o tamanho da amostra de bootstrap é escolhido para ser igual ao número de amostras no conjunto de treinamento original, o que geralmente fornece uma boa compensação de variação de polarização. Para o número de recursos d em cada divisão, queremos escolher um valor que seja menor do que o número total de recursos no conjunto de treinamento. Um padrão razoável que é usado no *scikit-learn* e outras implementações é $\mathbf{d} = \sqrt{\mathbf{m}}$, onde m é o número de recursos no conjunto de treinamento. Convenientemente, não precisamos construir o classificador de floresta aleatório a partir de árvores de decisão individuais por nós mesmos, porque já existe uma implementação no *scikit-learn*, Figura 13, que se pode usar:

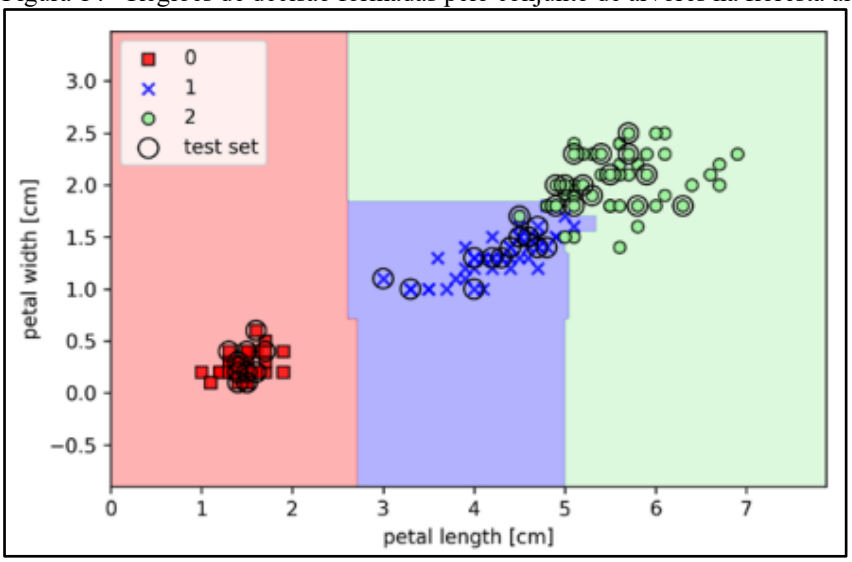
Figura 13 - Implementação no *Scikit-learn*

```
>>> from sklearn.ensemble import RandomForestClassifier
>>> forest = RandomForestClassifier(criterion='gini',
...                                 n_estimators=25,
...                                 random_state=1,
...                                 n_jobs=2)
>>> forest.fit(X_train, y_train)
>>> plot_decision_regions(X_combined, y_combined,
...                       classifier=forest, test_idx=range(105,150))
>>> plt.xlabel('petal length')
>>> plt.ylabel('petal width')
>>> plt.legend(loc='upper left')
>>> plt.show()
```

Fonte: RASCHKA; MIRJALILI (2017, p. 100).

Depois de executar o código anterior, devem-se ver as regiões de decisão formadas pelo conjunto de árvores na floresta aleatória, conforme mostrado na Figura 14:

Figura 14 - Regiões de decisão formadas pelo conjunto de árvores na floresta aleatória



Fonte: RASCHKA; MIRJALILI (2017, p.101).

Esta imagem apresenta usando o código anterior treina-se uma floresta aleatória a partir de 25 árvores de decisão por meio do parâmetro *n_estimators* e usa-se o critério de entropia como uma medida de impureza para dividir os nós. Embora esteja cultivando uma floresta aleatória muito pequena a partir de um conjunto de dados de treinamento muito pequeno, usa-se o parâmetro n_jobs para fins de demonstração, o que permite paralelizar o treinamento do modelo usando vários núcleos de nosso computador (aqui, dois núcleos) (RASCHKA; MIRJALILI, 2017).

2.5.4 Extra *Trees Classifier*

Muito semelhante ao *Random Forest*, o Extra *Trees* apresenta algumas diferenças e vantagens, dentre elas é o fato de um *Extra Tree Classifier* ser mais rápido em razão de não realizar alguns cálculos para separar as amostras entre as árvores. Ao invés disso, ele faz de forma randômica enquanto que o *Random Forest* tentará sempre otimizar a melhor separação.

Segundo ML (2021) uma árvore de classificação extremamente aleatória que escolhe as divisões de nó com a menor entropia entre um conjunto de k (dados por características máximas) pontos de divisão aleatórios, árvores extras são úteis em conjuntos como *Random Forest* ou *AdaBoost.* As forças das árvores extras em comparação com as árvores de decisão padrão são sua eficiência computacional e menor variação de previsão.

Nessa mesma linha de considerações Geurts, Ernst e Wehenkel (2006) enfatizam que o algoritmo *Extra-Trees* constrói um conjunto de árvores de decisão ou regressão não ajustadas de acordo com o procedimento clássico de cima para baixo, suas duas principais diferenças com outros métodos de conjunto baseados em árvores são que ele divide os nós escolhendo pontos de corte totalmente ao acaso e que usa todo o exemplo de aprendizagem (em vez de uma réplica *bootstrap*) para fazer as árvores crescer.

Figura 15 - Algoritmo *Extra-tree*

Table 1 Extra-Trees splitting algorithm (for numerical attributes)

Split_a_node(S)
Input: the local learning subset S corresponding to the node we want to split
Output: a split $[a < a_c]$ or nothing
– If **Stop_split**(S) is TRUE then return nothing.
– Otherwise select K attributes $\{a_1, \ldots, a_K\}$ among all non constant (in S) candidate attributes;
– Draw K splits $\{s_1, \ldots, s_K\}$, where $s_i = \textbf{Pick\_a\_random\_split}(S, a_i), \forall i = 1, \ldots, K$;
– Return a split s_* such that $\text{Score}(s_*, S) = \max_{i=1,\ldots,K} \text{Score}(s_i, S)$.

Pick_a_random_split(S,a)
Inputs: a subset S and an attribute a
Output: a split
– Let $a^S_{\max}$ and $a^S_{\min}$ denote the maximal and minimal value of a in S;
– Draw a random cut-point a_c uniformly in $[a^S_{\min}, a^S_{\max}]$;
– Return the split $[a < a_c]$.

Stop_split(S)
Input: a subset S
Output: a boolean
– If $|S| < n_{\min}$, then return TRUE;
– If all attributes are constant in S, then return TRUE;
– If the output is constant in S, then return TRUE;
– Otherwise, return FALSE.

Fonte: GEURTS; ERNST; WEHENKE (2006, p. 7).

O procedimento de divisão Extra-Árvores para atributos numéricos é fornecido na Tabela da Figura 15 tem dois parâmetros: K, o número de atributos selecionados

aleatoriamente em cada nó e ηmin e o tamanho mínimo da amostra para dividir um nó, sendo usadas várias vezes com a amostra de aprendizagem original (completa) para gerar um modelo de conjunto (denotamos por M o número de árvores desse conjunto). As previsões das árvores são agregadas para produzir a previsão final, por maioria de votos em problemas de classificação e média aritmética em problemas de regressão (GEURTS; ERNST; WEHENKEL, 2006).

Do ponto de vista do viés-variância, de acordo com Geurts, Ernst e Wehenkel (2006), a lógica por trás do método Extra-Árvores é que a aleatorização explícita do ponto de corte e do atributo combinada com a média do conjunto deve ser capaz de reduzir a variância mais fortemente do que os esquemas de randomização mais fracos usados por outros métodos. O uso da amostra de aprendizagem original completa, em vez de réplicas de *bootstrap*, é motivado para minimizar o viés. Já, do ponto de vista computacional, a complexidade do procedimento de crescimento das árvores é da ordem de N *log*, assumindo árvores balanceadas. N em relação ao tamanho da amostra de aprendizagem, como a maioria dos outros procedimentos de cultivo de árvores. No entanto, dada a simplicidade do procedimento de divisão do nó, espera-se que o fator constante seja muito menor do que em outros métodos baseados em conjunto que otimizam localmente os pontos de corte. Os parâmetros K, ηmine M têm efeitos diferentes, onde K determina a força do processo de seleção de atributos, ηmin a força do ruído de saída da média e M a força da redução da variância da agregação do modelo de conjunto, esses parâmetros podem ser adaptados às especificações do problema de forma manual ou automática.

No entanto, por preferência usam-se configurações padrão para eles, a fim de maximizar as vantagens computacionais e autonomia do método, a seção 3 estuda essas configurações padrão em termos de robustez e sub-otimização em vários contextos. Por fim, para especificar o valor do parâmetro principal K, usa-se a notação ETK, onde K é substituído por 'd' para dizer que as configurações padrão são usadas, por '$*$' para denotar os melhores resultados obtidos ao longo da faixa de valores possíveis de K, e por 'cv' se K for ajustado por validação cruzada (GEURTS; ERNST; WEHENKEL, 2006).

2.5.5 *Gradient Boosting Classifier*

Ao contrário dos conjuntos de média, os conjuntos de reforço constroem seus estimadores iterativamente. O conhecimento aprendido com o conjunto inicial é usado para

construir seus sucessores, mas essa é a principal desvantagem de impulsionar conjuntos, onde o paralelismo é inviável. Colocando o paralelismo de lado, essa natureza iterativa do conjunto exige que uma taxa de aprendizagem seja definida, isso ajuda o algoritmo de descida de gradiente a atingir os mínimos da função de perda facilmente. o exemplo abaixo utiliza 500 árvores, cada uma com no máximo 3 nós e uma taxa de aprendizado de 0,01. Além disso, a perda de mínimos quadrados Least Squares (LS), Figura 16, é usada também (AMR, 2020).

Figura 16 - *Least Squares* (LS)

```
from sklearn.ensemble import GradientBoostingRegressor

rgr = GradientBoostingRegressor (
  n_estimators=1000, learning_rate=0.01, max_depth=3, loss='ls'
)
rgr.fit(x_train, y_train)
y_test_pred = rgr.predict(x_test)
```

Fonte: AMR (2020, p. 235-236).

Este novo algoritmo apresenta o seguinte desempenho no conjunto de teste, Figura 17:

Figura 17 - Resultado Least Squares

```
# R2: 0.92, MSE: 3.93, RMSE: 1.98, MAE: 1.42
```

Fonte: AMR (2020, p. 235-236).

Observa-se que esta configuração, conforme Amr (2020), que se obteve um MSE menor em comparação com a floresta aleatória, enquanto a floresta aleatória teve um MAE melhor. Outra função de perda que o regressor de reforço de gradiente pode usar é *Least Absolute Deviation* (LAD).

LAD pode ajudar no tratamento com outliers, e às vezes, pode reduzir o desempenho do MAE do modelo no conjunto de teste. No entanto, não melhorou o MAE para o conjunto de dados em questão. Também tem uma perda percentual (AMR, 2020).

Neste contexto Amr (2020), enfatiza que os principais hiperparâmetros a serem definidos são o número de árvores, a profundidade das árvores, a taxa de aprendizagem e a

função de perda. Como regra geral, deve-se ter como objetivo um número maior de árvores e uma taxa de aprendizado baixa, esses dois hiperparâmetros são inversamente proporcionais um ao outro. O controle da profundidade de suas árvores depende puramente de seus dados, em geral, precisam-se ter árvores rasas e deixar o aumento de empoderá-las. No entanto, a profundidade da árvore controla o número de interações de recursos que se quer capturar. Em um *stub* (uma árvore com uma única divisão), apenas um recurso pode ser aprendido por vez, uma árvore mais profunda se assemelha a uma condição *if* aninhada em que alguns recursos a mais estão em jogo a cada vez, onde normalmente começo com *max_depth* definido para cerca de 3 e 5 e ajustado ao longo do caminho.

2.5.6 *XG Boost*

Para Malik, Harode e Kunwar (2020), *XG Boost* é um modelo que utiliza vários conceitos de modelos existentes e aprimorados para obter uma melhor performance combinando várias pequenas árvores de eliminação onde ele utiliza uma técnica do gradiente descendente para otimizá-las.

Aumento de gradiente é um caso especial de algoritmo de reforço onde os erros são minimizados por um algoritmo de descida de gradiente e produzem um modelo na forma de modelos de previsão fracos, por exemplo, árvores de decisão. A principal diferença entre boosting e gradiente boosting é como ambos os algoritmos atualizam o modelo alunos fracos a partir de previsões erradas, um aluno fraco é um classificador que tem uma correlação fraca com o valor real. O aumento de gradiente ajusta os pesos pelo uso de gradiente uma direção na função de perda usando um algoritmo chamado gradiente descendente, que otimiza iterativamente a perda do modelo atualizando os pesos. Perda normalmente significa a diferença entre o valor previsto e o valor real. Para algoritmos de regressão, usamos a perda Mean Squared Error (MSE), Figura 18, enquanto para problemas de classificação usamos a perda logarítmica (Malik; Harode; Kunwar, 2020).

Figura 18 - *Mean Squared Error (MSE)*

$$w = w - \eta \nabla w$$
$$\nabla w = \frac{\partial L}{\partial w} \quad where\ L\ is\ loss$$

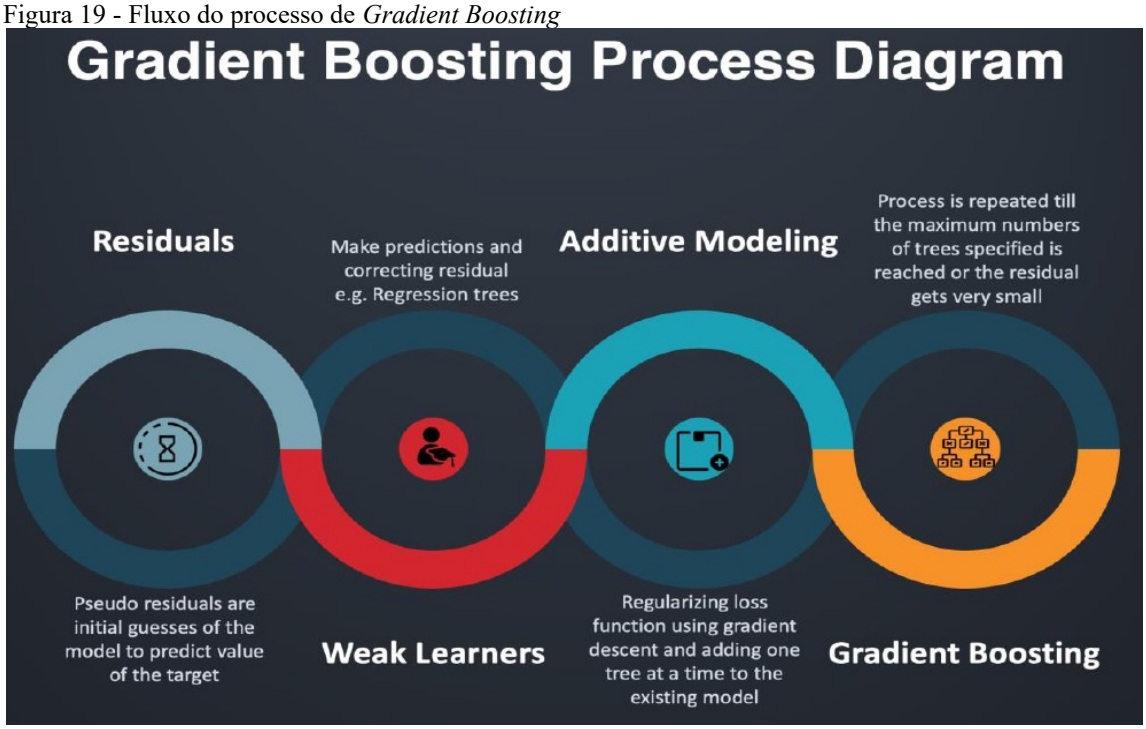
Fonte: MALIK; HARODE; KUNWAR (2020, p. 10).

Onde: w representa o vetor de peso, η é a taxa de aprendizagem.

2.5.1.4. Processo de aumento de gradiente

A Figura 19 representa o fluxo do processo de *Gradient Boosting* descrito por Malik, Harode e Kunwar (2020).

Figura 19 - Fluxo do processo de *Gradient Boosting*



Fonte: (MALIK; HARODE; KUNWAR, 2020, p.10).

2.5.1.5. Fluxo do processo de Gradient Boosting

O aumento de gradiente usa a modelagem aditiva na qual uma nova árvore de decisão é adicionada uma de cada vez a um modelo que minimiza a perda usando a descida de gradiente. As árvores existentes no modelo permanecem intocadas e, portanto, reduzem a taxa de sobreajuste. A saída da nova árvore é combinada com a saída das árvores existentes até que a perda seja minimizada abaixo de um limite ou o limite especificado de árvores seja alcançado (MALIK; HARODE; KUNWAR, 2020).

Para Malik, Harode e Kunwar (2020) a Modelagem Aditiva em matemática é a divisão de uma função na adição de N subfunções. Em termos estatísticos, pode ser pensado como um modelo de regressão no qual a resposta y é a soma aritmética dos efeitos individuais das variáveis preditoras x.

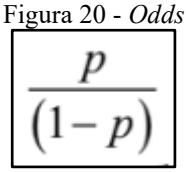
2.5.7 *Logistic Regression*

O ponto principal do *Logistic Regression* é estimar a probabilidade de uma amostra pertencer a um grupo comparando com a probabilidade de pertencer às demais. Por exemplo, em uma classificação binária, aprovado ou reprovado, caso exista uma chance de 75% da amostra pertencer a classe aprovada, logo, a chance de ele pertencer a classe reprovado será de 25% (RASCHKA; MIRJALILI, 2017).

Conforme Raschka e Mirjalili (2017), A regressão logística é um modelo de classificação fácil de implementar, mas funciona muito bem em classes linearmente separáveis. É um dos algoritmos mais usados para classificação na indústria, o modelo de regressão logística também é um modelo linear para classificação binária que pode ser estendido para classificação multiclasse.

A Figura 20 ilustra um exemplo para um melhor entendimento da ideia por trás da regressão logística como um modelo probabilístico, primeiro é introduzido a razão de chances: as chances a favor de um evento particular. O *odds* razão pode ser escrito como:

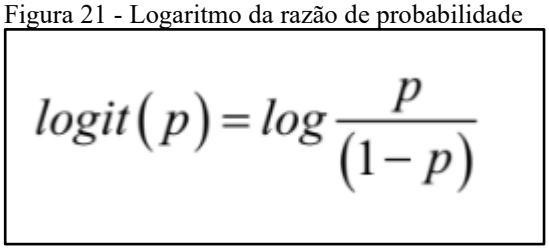
Figura 20 - *Odds* razão



Fonte: RASCHKA; MIRJALILI (2017, p. 59).

Onde p representa a probabilidade do evento positivo. O termo evento positivo não significa necessariamente bom, mas se refere ao evento que queremos prever, como por exemplo, a probabilidade de um paciente ter uma determinada doença, pode-se pensar no evento positivo como rótulo de classe y = 1. Então definir a função *logit*, que é simplesmente o logaritmo da razão de probabilidade (*log-odds),* Figura 21*:*

Figura 21 - Logaritmo da razão de probabilidade



Fonte: RASCHKA; MIRJALILI (2017, p. 59).

Log se refere ao logaritmo natural, pois é uma convenção comum na ciência da computação. A função logit leva como valores de entrada no intervalo de 0 a 1 e os transforma em valores ao longo de todo o intervalo de número real que podemos usar para expressar uma relação linear entre os valores de característica e as probabilidades de *log*.

2.6 TREINAMENTO E OTIMIZAÇÃO DE ALGORITMOS

Conforme mencionado anteriormente, o processo de treinamento é uma das formas pela qual se dá o aprendizado de máquina. Existem várias técnicas que podem ser aplicadas junto com uma boa escolha de métricas para se monitorar a assertividade e eficácia do modelo.

2.6.1 *Grid Search*

O *Grid Search* CV é uma forma de iterar entre várias combinações de parâmetros e realizar o treinamento de acordo com essas soluções e ver o que foi o melhor resultado obtido com base nas métricas definidas. *Grid Search* CV é um método de *loop* em todas as combinações de hiperparâmetros possíveis e empregando validação cruzada para escolher os hiperparâmetros ideais. Para cada combinação de hiperparâmetros, não se pode limitar a apenas uma pontuação de precisão (AMR, 2020).

Portanto, segundo Amr (2020), para obter uma melhor compreensão da precisão do estimador de cada combinação é usada a validação cruzada K-*fold*. Posteriormente os dados são divididos em várias dobras e para cada iteração todas as dobras exceto uma é usada para treinamento e a restante é usada para teste. Este método de ajuste de hiperparâmetros realiza uma pesquisa exaustiva sobre todas as combinações de parâmetros possíveis, surgindo daí o prefixo Grid.

De acordo com Mujtaba (2020), *Grid Search* CV ou pesquisa em grade é o processo de realizar o ajuste de hiperparâmetros para determinar os valores ideais para um determinado modelo. O desempenho de um modelo depende significativamente do valor dos hiperparâmetros, visto que não há como saber com antecedência os melhores valores para hiperparâmetros, então é necessário tentar todos os valores possíveis para saber os valores ideais. Fazer isso manualmente pode levar uma quantidade considerável de tempo e recursos e, portanto, usa-se *Grid Search* CV para automatizar o ajuste de hiperparâmetros. Sendo

assim, fica claro que o custo de se treinar utilizando vários parâmetros pode até ser demorado, mas os resultados são apresentados como melhores modelos.

2.6.2 Método *K-Folding*

Uma analogia a um jogo de cartas, o *k-folding,* como descreve Izenman (2008), é para evitar que por sorte sejam sorteadas apenas cartas boas ou ruins para um jogador e isso tenha um impacto no resultado final. Logo é como se durante o treinamento do modelo fossem "dadas as cartas" para treinamento e teste múltiplas vezes. Quanto mais vezes esse processo for repetido, mais se garante que não é uma aleatoriedade o fato do modelo ser bom ou ruim e sim o fato de que ele aprendeu os padrões existentes nos dados.

Conforme Izenman (2008), *V - Fold Cross-validation* ou Validação cruzada em V: Divida aleatoriamente todo o conjunto de dados em V grupos não sobrepostos de tamanho aproximadamente igual; remova um dos grupos e ajuste o modelo usando os dados combinados dos outros grupos V −1 (que formam o conjunto de aprendizagem); use o grupo omitido como o conjunto de teste, preveja seus valores de saída usando o modelo ajustado e calcule o erro de previsão para o grupo omitido; repita este procedimento V vezes, removendo cada vez um grupo diferente; em seguida, calcule a média dos erros de predição V resultantes para estimar o erro de teste. O número de grupos V pode ser qualquer número de 2 ao tamanho da amostra.

2.7 MÉTRICAS DE PERFORMANCE DE ALGORITMO

As métricas de performance ajudam a identificar se o modelo atende aos requisitos estabelecidos, isto é, soluciona ou não o problema estabelecido. Elas também permitem verificar problemas como *overfitting e underfitting* (RASCHKA; MIRJALILI, 2017).

2.7.1 Acurácia

Uma das métricas mais básicas para se avaliar os modelos e que não pode ser avaliada de forma isolada, comumente usada é a precisão da classificação, que é definida como a proporção de instâncias classificadas corretamente (RASCHKA; MIRJALILI, 2017).

Para Raschka e Mirjalili (2017), supondo que existem 10 amostras para serem classificadas entre pacientes que desenvolveram infecção ou não, caso o modelo acerte 9 das

10, terá uma acurácia de 90%. Porém, nos cenários onde existe naturalmente nos dados uma probabilidade maior de uma classe, a acurácia se torna insuficiente. Por exemplo, supondo que apenas um paciente de cada dez desenvolva uma infecção. Se o modelo simplesmente chutar que ninguém irá desenvolver infecção ele irá acertar naturalmente 90% das vezes. Assim, outras métricas como o Recall tornam-se essenciais.

2.7.2 *Recall*

Para Hackeling (2017), a precisão é a fração das previsões positivas corretas. No classificador de *spam* por SMS, precisão é a fração das mensagens classificadas como spam que na verdade é *spam*. Às vezes chamada de sensibilidade em domínios médicos, a lembrança é a fração dos casos verdadeiramente positivos que o classificador reconheceu. Uma pontuação de recordação de 1 indica que o classificador não fez nenhuma previsão falsa negativa. Para o classificador de spam por SMS, *recall* é a fração das mensagens verdadeiramente spam que foram classificadas como *spam*.

Individualmente, precisão e *recall* raramente são informativos; ambas são visões incompletas do desempenho de um classificador. Tanto a precisão quanto a recuperação podem falhar em distinguir classificadores com bom desempenho de certos tipos de classificadores com desempenho insatisfatório (HACKELING, 2017).

Hackeling (2017) ainda destaca que, um classificador trivial poderia facilmente atingir uma pontuação de recordação perfeita, prevendo positivo para cada instância. Por exemplo, supondo que um conjunto de testes contém 10 exemplos positivos e 10 exemplos negativos, um classificador que prevê positivo para todos os exemplos alcançará um *recall* de 1 (um). Um classificador que prevê negativo para todos os exemplos, ou um que faz apenas previsões falsas positivas e negativas verdadeiras, alcançará uma pontuação de recall de 0 (zero). Da mesma forma, um classificador que prevê que apenas uma única instância seja positiva e correta atingirá a precisão perfeita. A Figura 22 exemplifica um classificador.

Figura 22 - Precisão e *Recall*

```
# In[2]:
precisions = cross_val_score(classifier, X_train, y_train, cv=5,
  scoring='precision')
print('Precision: %s' % np.mean(precisions))
recalls = cross_val_score(classifier, X_train, y_train, cv=5,
  scoring='recall')
print('Recall: %s' % np.mean(recalls))

# Out[2]:
Precision: 0.992542742398
Recall: 0.683605030275
```

Fonte: HACKELING (2017, p. 99).

A precisão do classificador utilizado no exemplo é 0.992; quase todas as mensagens que previu como *spam* era na verdade *spam*. Seu *recall* é menor, indicando que classificou incorretamente aproximadamente 32% das mensagens de spam como *spam* (HACKELING, 2017).

2.7.3 *F1 Score*

Para Hackeling (2017), a medida F1 é a média harmônica das pontuações de precisão e *recall,* ela penaliza classificadores com precisão desequilibrada e pontuações de *recall*, como o classificador trivial que sempre prevê a classe positiva. Um modelo com precisão perfeita e pontuações de *recall* atingirá uma pontuação F1 de 1, conforme exemplificado na Figura 23.

Figura 23 - Calculando a medida de F1

```
# In[3]:
f1s = cross_val_score(classifier, X_train, y_train, cv=5,
  scoring='f1')
print('F1 score: %s' % np.mean(f1s))

# Out[3]:
F1 score: 0.809067846627
```

Fonte: HACKELING (2017, p. 99).

Os modelos são avaliados usando os escores F0, 5 e F2, que influenciam a precisão em relação à recuperação e a recuperação em relação à precisão, respectivamente.

2.7.4 AUC e ROC Curves (área sob a curva e curvas ROC)

Para Hackeling (2017), uma curva *Receiver Operating Characteristic* (ROC), visualiza o desempenho de um classificador. Diferente da precisão, a curva ROC é insensível a conjuntos de dados com proporções de classe desequilibradas; ao contrário da precisão e da recuperação, a curva ROC ilustra o desempenho do classificador para todos os valores do limite de discriminação. As curvas ROC traçam o classificador recall contra sua precipitação. *Fall-Out*, ou taxa de falsos positivos, é o número de falsos positivos divididos pelo número total de negativos. É calculado usando a seguinte fórmula, conforme

Figura 24:

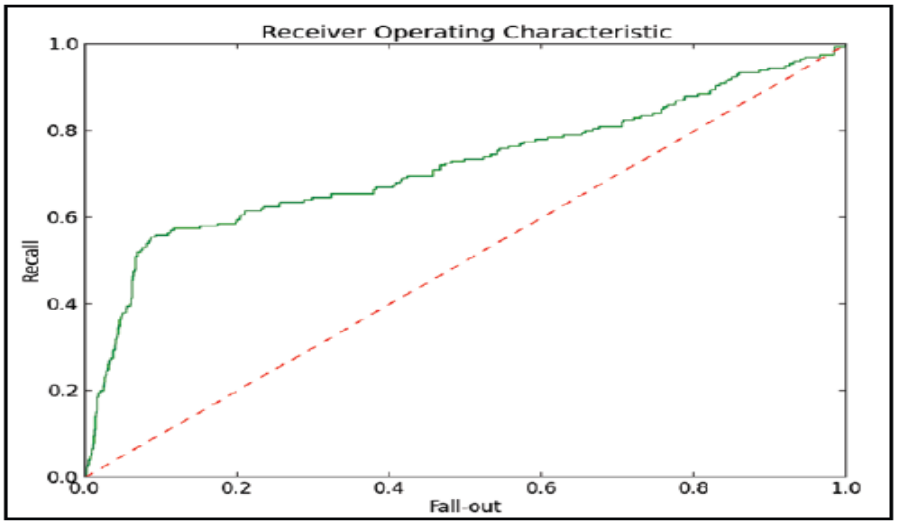
Figura 24 - Fórmula de ROC

$$F = \frac{FP}{TN + FP}$$

Fonte: HACKELING (2017, p. 100).

Para Hackeling (2017), AUC é a área sob a curva ROC; este reduz a curva ROC a um único valor que representa o desempenho esperado do classificador. A linha tracejada na figura a seguir é para um classificador que prevê classes aleatoriamente; tem uma AUC de 0,5. A curva sólida é para um classificador que supera a suposição aleatória, como ilustra a Figura 25:

Figura 25 - A curva de ROC



Fonte: HACKELING (2017, p. 100).

Plotando uma curva ROC para o classificador de *spam SMS*, Figura 26:

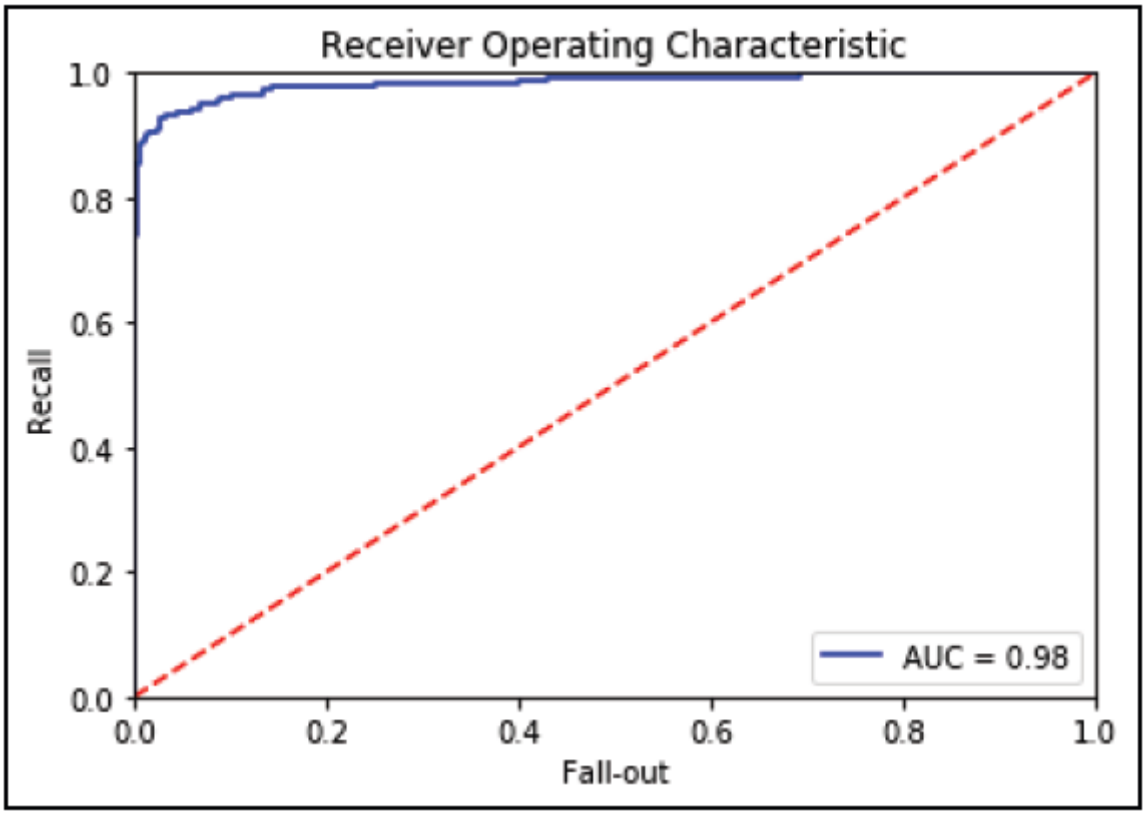
Figura 26 - Plotando uma curva ROC

```
# In[5]:
predictions = classifier.predict_proba(X_test)
false_positive_rate, recall, thresholds = roc_curve(y_test,
  predictions[:, 1])
roc_auc = auc(false_positive_rate, recall)
plt.title('Receiver Operating Characteristic')
plt.plot(false_positive_rate, recall, 'b', label='AUC = %0.2f' %
  roc_auc)
plt.legend(loc='lower right')
plt.plot([0, 1], [0, 1], 'r--')
plt.xlim([0.0, 1.0])
plt.ylim([0.0, 1.0])
plt.ylabel('Recall')
plt.xlabel('Fall-out')
plt.show()
```

Fonte: HACKELING (2017, p. 101).

A partir do gráfico ROC AUC, é aparente que o classificador supera a suposição aleatória; a maior parte da área do gráfico encontra-se sob sua curva, Figura 27:

Figura 27 - Gráfico ROC AUC



Fonte: HACKELING (2017, p. 101).

2.7.5 Matriz de confusão

Para Han, Kamber e Pei (2012), matriz de confusão é uma forma simplificada (tabular) de identificar a performance do modelo. Ela faz uso do resultado das classificações que também é utilizado para calcular as métricas abordadas anteriormente.

A matriz de confusão é uma ferramenta útil para analisar quão bem seu classificador pode reconhecer tuplas de diferentes classes. TP e TN nos dizem quando o classificador está fazendo as coisas certas, enquanto FP e FN dizem quando o classificador está fazendo coisas erradas, ou seja, etiquetagem incorreta.

Ainda segundo Han, Kamber e Pei (2012) dadas m classes (onde m ≥ 2), uma matriz de confusão é uma tabela de pelo menos tamanho m por m. Uma entrada, CM i, j nas primeiras m linhas em colunas indica o número de tuplas da classe i que foram rotuladas pelo classificador como classe j. Para um classificador ter boa precisão, idealmente a maioria das tuplas seria representada ao longo da diagonal da matriz de confusão, da entrada CM 1,1 à entrada CM m, m, com o restante das entradas sendo zero ou próximo de zero. Ou seja, idealmente, FP e FN estão em torno de zero. A Figura 28 demonstra a matriz de confusão:

Figura 28 - Matriz de confusão, mostrada com totais para tuplas positivas e negativas

		Predicted class		
		yes	*no*	Total
Actual class	*yes*	*TP*	*FN*	*P*
	no	*FP*	*TN*	*N*
	Total	P'	N'	$P + N$

Fonte: HAN; KAMBER; PEI (2012, p. 366).

2.8 TRABALHOS CORRELATOS

Samwald *et al.* (2012), A sintaxe *Arden* é uma linguagem de marcação usada para representar e compartilhar conhecimentos médicos. Essa linguagem do conhecimento clínico e científico é usada em um formato executável pelos sistemas de suporte a decisões clínicas para gerar alertas, interpretações e rastrear e gerenciar mensagens para os médicos. Essa sintaxe é usada para compartilhar conhecimento médico dentro e entre muitas instituições de

serviços de saúde. Os conjuntos de regras, chamados *Medical Logic Modules*, compreendem lógica suficiente para tomar uma única decisão médica. Os módulos de lógica médica são escritos na sintaxe *Arden* e são chamados por um programa, um monitor de eventos quando ocorre a condição na qual eles foram gravados para ajudar. Os autores ainda complementam que, a sintaxe da *Arden* era anteriormente um padrão da ASTM, publicado em 1992, e agora faz parte do *Health Level Seven International*. A sintaxe *Arden* versão 2.0 foi publicada pela HL7 em 1999. A sintaxe *Arden* versão 2.10 é a versão atual.

Criado um ambiente de desenvolvimento pronto para produção, compilador, regra, mecanismo e servidor de aplicativos para *Arden Syntax*. Ao longo de vários anos foi aplicado esse Sistema CDS baseado em sintaxe *Arden* em uma ampla variedade de domínios de problemas clínicos como hepatite, interpretação sorológica, monitoramento de infecções ou predição de eventos metastáticos em pacientes com melanoma. Descoberto que o padrão *Arden Syntax* é muito adequado para a implementação prática instalação de sistemas CDS. Entre as vantagens da *Arden Syntax*, está o status de HL7 ativamente desenvolvido padrão, a legibilidade da sintaxe e vários recursos sintáticos, como manipulação flexível de lista (SAMWALD *et al*., 2012).

Além disso, como enfatizam Samwald *et al.* (2012), um grande desafio encontrado foi a integração técnica dos sistemas CDS em sistemas heterogêneos existentes, sistemas de informação sanitária. Para resolver esse problema, foi trabalhado na incorporação do GLLO padrão HL7, que fornece uma interface padronizada e linguagem de consulta para acessar dados em sistemas de informação em saúde. Desta forma, a tomada de decisão na prática médica moderna é baseada em conhecimento médico cada vez mais complexo e evidência clínica. Isso dificulta a prestação do melhor atendimento possível nos ambientes normalmente encontrados nas configurações de assistência médica. Bem como, foi demonstrado que os sistemas de suporte à decisão clínica (CDS) podem melhorar significativamente a qualidade do tratamento caso atendam a critérios de projeto.

Arden Syntax é um padrão amplamente reconhecido para representar conhecimento clínico e científico em um formato executável que pode ser usado por esses sistemas CDS. Qualquer linguagem de programação comum pode, em teoria, ser usada para implementar sistemas CDS (SAMWALD *et al*., 2012).

Entretanto, o padrão *Arden Syntax* foi projetado para esse fim específico e está equipado com um conjunto de estruturas que o tornam especialmente útil para esta tarefa. A sintaxe Arden pode ser usada de uma maneira que faça com que o código do programa se

pareça com a linguagem natural, que por sua vez, facilita a compreensão do código por não especialistas em ciência da computação. Ele também possui uma escolha de dados tipos adaptados às necessidades da documentação médica, incluindo medidas de tempo e duração (SAMWALD *et al*., 2012)..

Além disso, como destacam Samwald *et al.* (2012), o uso da *Arden Syntax* torna possível representar independentemente das linguagens de programação e detalhes da informação escolhidos para uma Hospital Information System ou em português Sistemas de Informação do Hospital (HIS), facilitando a troca da lógica do CDS entre diferentes sistemas em diferentes locais. Neste sentido, a sintaxe Arden pode ser vista como um híbrido entre a produção clássica regras de representação e representação processual de algoritmos clínicos. O código é organizado em arquivos independentes chamados *Medical Logic Mod-(MLMs*). A execução de um MLMs pode ser desencadeada por dados ou eventos baseados em tempo ou por uma chamada direta.

Conforme ressaltam Samwald *et al.* (2012), o padrão Arden Syntax agora tem um histórico de cerca de duas décadas. Embora problemas com interoperabilidade sintática e semântica sistemas heterogêneos de informação em saúde tiveram um impacto negativo sobre a ampla adoção dessa norma no passado, recentes desenvolvimentos da norma são obrigados a minimizar esses problemas. Esses desenvolvimentos provavelmente aumentarão a utilização de padrão e pode reviver os ideais originais que impulsionam seu desenvolvimento: criação e compartilhamento gratuitos e sem ônus para melhorar a prática clínica e finalmente a qualidade de vida dos pacientes.

No que diz respeito às infecções hospitalares, segundo Lamma *et al.* (2006), estas são perigosas porque são causadas por bactérias que apresentam resistência aos antibióticos. Este problema é muito sério em todo o mundo. Na Itália, quase 8% dos pacientes internados em hospitais desenvolvem esse tipo de infecção. Para reduzir esse número políticas de controle de infecções devem ser adotadas.

Entre as tecnologias que visam auxiliar neste sentido, o MERCURIO é um pacote de software comercial que é o resultado da engenharia de um sistema de pesquisa desenvolvido no escopo do projeto TDMIN, uma empresa italiana de tecnologia da informação que atua no mercado de serviços de saúde e o Departamento de Eletrônica, Informática e Sistemática (DEIS) da Universidade de Bolonha. Os objetivos deste sistema são a validação de dados microbiológicos e a criação de um sistema de informações epidemiológicas em tempo real (LAMMA *et al.*, (2006). Conforme afirmam os autores, o sistema é útil para médicos de

laboratório, pois os auxilia na execução dos testes microbiológicos para os médicos, porque os apoia na definição da antibioticoterapia mais adequada e no monitoramento das infecções dos pacientes e para epidemiologistas porque permite identificar surtos e estudar infecções.

Contudo, para atingir esses objetivos Lamma *et al.* (2006) adotaram técnicas de inteligência artificial e em particular sistemas e dados especializados técnicas de mineração. Uma abordagem de sistema especialista foi aplicada a fim de obter clareza na possibilidade de explicar em detalhes as respostas dadas, flexibilidade a possibilidade de atualizar facilmente a base de conhecimento e confiabilidade a exatidão das respostas dadas.

Como resultado da atividade do projeto foi construído um sistema composto principalmente por um banco de dados epidemiológico projetado para armazenar dados epidemiológicos em um sistema baseado em conhecimento, chamado ESMIS, para validação e monitoramento em tempo real de um módulo estatístico para realizar análises estatísticas e identificar surtos. Os objetivos foram atingidos com a aplicação de um sistema especialista, mineração de dados e técnicas estatísticas (LAMMA *et al.*, (2006).

Para Lamma *et al.* (2006), O ESMIS provou ser útil para os médicos de laboratório, apoiando-os na execução de análises microbiológicas e ajudando-os a evitar erros perigosos causados por humanos e instrumentos. Em um estudo de teste realizado com dados de seis meses, o ESMIS alcançou uma precisão e especificidade muito altas nas tarefas de validação e relatório inteligente. Em relação à tarefa de teste do *rack* a sensibilidade foi muito boa.

Lamma *et al.* (2006), ainda completam que as técnicas de mineração de dados foram aplicados para descobrir automaticamente regras de associação a partir de dados microbiológicos e obter regras de alarme com eles. A abordagem de descoberta de conhecimento que seguimos provou ser muito eficaz na validação de parte da base de conhecimento do ESMIS (escrita de acordo com as diretrizes da National Committee for Clinical Laboratory Standards (NCCLS) e com especialistas em microbiologia), e também na extensão com novas regras de validação confirmadas pelos microbiologistas entrevistados e específicos para o laboratório hospitalar considerado.

A contribuição dessa abordagem, portanto, é dupla, conforme enfatizam Lamma *et al*. (2006):

- Fornece uma maneira de conscientizar os especialistas humanos sobre novas correlações entre alguns resultados de testes antimicrobianos que não foram notados anteriormente;

- Pode ser considerado um método automático para validar e possivelmente ampliar a base de conhecimento de um sistema especialista no domínio microbiológico. Nos experimentos realizados com essa abordagem em algumas espécies de bactérias, foi obtido resultados interessantes.

Com referência a resistência antimicrobiana, de acordo com Baysari *et al.* (2016), representa hoje uma das maiores ameaças à saúde humana, com o temor de uma era pós-antibiótica. Com o desenvolvimento de resistência generalizada contra praticamente todos os antimicrobianos disponíveis e poucos novos antimicrobianos na tubulação é agora necessário melhorar o uso de antimicrobianos em ambientes de atenção aguda e primária para a contenção da resistência. Estudos demonstraram que o uso de antimicrobianos em hospitais geralmente é inadequado, às vezes até 50% das vezes. O termo "uso antimicrobiano inadequado" é usado para descrever uma variedade de casos, por exemplo, antimicrobianos sendo prescritos quando não são necessários, antimicrobianos sendo prescritos em doses inadequadas, ou a seleção de um antimicrobiano de amplo espectro quando está disponível um antimicrobiano de espectro estreito mais apropriado.

Ainda segundo Baysari *et al.* (2016), as intervenções de Tecnologia da Informação (TI) para apoiar a prescrição antimicrobiana apropriada apresentam uma nova e empolgante perspectiva para enfrentar esse desafio internacional. À medida que os hospitais avançam na adoção de sistemas eletrônicos de prontuários médicos surgem novas oportunidades para a integração de políticas hospitalares antimicrobianas, apoio à decisão e uso e vigilância de antimicrobianos nos sistemas eletrônicos. Várias tentativas foram feitas para compilar as evidências para alcançar uma melhor prescrição antimicrobiana com suporte a decisões computadorizadas, mas há análises limitadas da eficácia das intervenções de TI em geral. Quando um hospital está contemplando a implementação de soluções de TI para direcionar prescrições antimicrobianas inapropriadas, há pouca informação disponível sobre a forma ou o tipo de intervenção em TI mais adequado para atingir esse objetivo. O objetivo desta revisão sistemática foi revisar as evidências da eficácia das intervenções de TI na melhoria da prescrição microbiológica. Também foi realizado uma metanálise, cujo objetivo foi determinar o impacto da implementação de intervenções de TI em três medidas de resultado: uso adequado de antimicrobianos, mortalidade de pacientes e tempo de permanência no hospital.

Neste sentido, a evidência do impacto das intervenções de TI nos resultados de saúde, como mortalidade e tempo de permanência, é mais variável. A variedade de projetos de estudo e medidas de resultado usadas para avaliar as intervenções de TI impediram comparações significativas entre os diferentes tipos de sistemas de TI. Atualmente, poucas evidências estão disponíveis para as organizações auxiliares na seleção informada de soluções de TI para a prescrição de antimicrobianos. Isso resultará inevitavelmente em muitas seleções serem conduzidas por relatos anedóticos de utilidade e facilidade de uso do sistema (BAYSARI *et al*., 2016).

Na opinião de Adlassnig, Blacky e Koller (2009), a crescente disponibilidade de dados médicos digitalizados de pacientes em um hospital permite a identificação e o monitoramento abrangentes de infecções hospitalares. Os sistemas de informação agora rotineiramente usados em hospitais são um dos fundamentos básicos desse procedimento. Os sistemas são capazes de armazenar, transferir e recuperar um corpo cada vez maior de dados digitalizados sobre o histórico médico dos pacientes, o resultado do exame físico, os diferentes resultados dos testes de laboratório e os resultados de várias investigações clínicas. Esses sistemas são conhecidos como Hospital Information Systems ou em português Sistemas de Informações Hospitalares (HISs), cujas muitas funções incluem a administração de dados referentes à admissão, transferência e alta de pacientes a fim de tornar esses dados acessíveis aos Medical Information Systems ou em português Sistemas de Informações Médicas (MISs) nas diferentes enfermarias e fora departamentos de pacientes que contêm os dados médicos dos pacientes como sistemas de informações de laboratório (LISs) com os resultados laboratoriais obtidos respectivamente bem como Sistemas de Gerenciamento de Dados de Pacientes (PDMSs) nas UTIs com dados clínicos, laboratoriais, baseados em equipamentos e de enfermagem.

Além disso, o suporte baseado em conhecimento de alta qualidade para a tomada de decisões médicas com base nesses dados do paciente armazenados nos respectivos sistemas de informação requer que o conhecimento médico seja representado de maneira formal e armazenado em um sistema de computador. Isso pode ocorrer na forma de interpretações de achados laboratoriais raros ou complexos ou de definições de sintomas, doenças e processos de tratamento em computador e suas inter-relações, ou regras ou formas tabuladas de procedimentos médicos para tomada de decisão (ADLASSNIG; BLACKY; KOLLER, 2009).

Convém destacar, conforme enfatizam Adlassnig, Blacky e Koller (2009), que os avanços nos métodos de representação formal e processamento do conhecimento médico

alcançados nos campos da inteligência artificial, teoria dos conjuntos nebulosos e lógica nebulosa permitem o processamento computadorizado do conhecimento médico originalmente disponível em linguagem natural.

Adlassnig, Blacky e Koller (2009) consideram que ao aplicar métodos de inteligência artificial e teoria fuzzy, o programa de identificação e monitoramento existente *Moni / Surveillance* foi equipado com inteligência baseada no conhecimento que executa etapas analíticas complexas automaticamente as substancia e as torna compreensíveis e reproduzíveis. Desta forma, acredita-se que a aplicação rotineira deste programa fará uma contribuição significativa para a gestão da qualidade no Hospital Geral de Viena. Em particular ajudará os médicos no tratamento a reduzir a taxa de infecção hospitalar nas UTIs e, portanto, pode potencialmente servir como uma medida significativa de redução de custos.

Como identificam Rubin *et al.* (2008), a divulgação pública de taxas de infecções hospitalares (HAI) está ganhando força nos Estados Unidos, já que funcionários públicos procuram proporcionar aos consumidores de saúde a capacidade de tomar decisões mais bem informadas sobre seus cuidados. A geração e interpretação das taxas de Infecções Relacionadas a Assistência à Saúde, no entanto, é complexa. A vigilância tradicional de HAI, liderada por Profissionais de Controle de Infecção (ICPs) é um processo trabalhoso que envolve critérios de caso altamente subjetivos que não são verdadeiros "padrões-ouro". A aplicação desses critérios também pode ser altamente influenciada por fatores como treinamento e experiência em ICP, práticas hospitalares individuais e características específicas do hospital. Isso levanta sérias preocupações sobre a confiabilidade desses critérios e as taxas calculadas por diferentes indivíduos que aplicam os critérios aos seus dados. Como tal, a interpretação precisa desses dados é implausível e seu uso pelo público para comparar o desempenho de diferentes instalações pode ser inapropriado e mais confuso do que útil.

Como descrito por Rubin *et al.* (2008), métodos alternativos e simplificados de vigilância que dependem inteiramente de critérios objetivos, como dados de microbiologia extraídos de um prontuário eletrônico, apresentam uma opção atraente devido à sua capacidade de serem automatizados e porque sua confiabilidade é potencialmente muito maior. Esses métodos podem não ser tão precisos na estimativa de verdadeiras taxas de infecção, no entanto, porque se baseiam em um conjunto menor de dados e como resultado são menos seletivos. Ainda assim um sistema objetivo e muito mais confiável aplicado em

diferentes instituições pode ser mais útil quando o objetivo final é a comparação de taxas entre essas instituições.

Vale ressaltar que, como não existe um "padrão-ouro" verdadeiro para o diagnóstico de CRBSI, é difícil avaliar e comparar essas duas abordagens de vigilância usando dados do mundo real. Em vez disso, optou-se por criar um modelo de simulação baseado em agente para simular a ocorrência e a vigilância do CRBSI, pois essa abordagem nos permitiu saber definitivamente quais pacientes tinham infecções verdadeiras. Em seguida foi explorada a troca entre confiabilidade e validade ao usar os dois métodos de vigilância para estimar a verdadeira taxa de CRBSI de uma instituição e comparar a classificação das taxas estimadas em várias instituições (RUBIN *et al*., 2008).

Sendo assim, como alegam Rubin *et al.* (2008), com o atual esforço nacional em direção à notificação pública obrigatória das taxas de IRAS maior atenção deve ser dada aos métodos utilizados para a vigilância de infecções e à maneira inconsistente com que os critérios de caso são aplicados em diferentes instituições. As descobertas desta pesquisa sugerem que um conjunto mais simplificado de critérios de vigilância pode melhorar a confiabilidade e a comparabilidade das estimativas de taxas de IRAS nas instituições, apesar do fato de que esses critérios podem produzir estimativas de taxas geralmente menos precisas no nível individual das instalações.

O estudo realizado por Bagci *et al.* (2012) apresentou um novo sistema de Detecção Assistida por Computador (CAD) para detectar e quantificar automaticamente opacidades anormais de ramificação nodular em Tomografia Computadorizada de tórax (TC), denominadas Opacidades de Árvore em Brotamento (TIB) pela literatura radiológica.

Neste sentido, doenças pulmonares infecciosas, como a nova gripe H1N1 de origem suína, tuberculose estão entre as principais causas de incapacidade e morte em todo o mundo. O exame de Tomografia Computadorizada dos pulmões durante infecções agudas do trato respiratório tornou-se uma parte importante do atendimento ao paciente tanto no diagnóstico quanto no monitoramento da progressão ou resposta à terapia. Embora o diagnóstico correto do padrão de TIB seja muito importante, também é uma das tarefas mais difíceis para os radiologistas, porque o contraste das lesões é geralmente baixo e os padrões da doença são muito complexos. Todas essas limitações sugerem que a DAC pode dar uma contribuição valiosa ao tratamento de infecções do trato respiratório. (BAGCI *et al.*, 2012)

Assim como, como ressaltam Bagci *et al.* (2012), também foi comparado a pontuação computacional do sistema CAD proposto com a classificação visual subjetiva. É

obtida uma alta correlação entre as pontuações objetiva (CAD) e subjetiva (classificação visual), o que implica em uma precisão altamente satisfatória do sistema CAD proposto.

Para Warner *et al.* (2016) As Complicações Adquiridas em Hospitais são problemas sérios que afetam as instituições de saúde modernas. Estima-se que os HACs resultem em um aumento de aproximadamente 10% no total de custos hospitalares em hospitais nos Estados Unidos da América (*EUA)*. Com os gastos hospitalares dos EUA totalizando quase US $ 900 bilhões por ano, os danos causados pelos HACs não são um problema pequeno. A detecção e prevenção precoces de HACs podem reduzir significativamente as tensões no sistema de saúde dos *EUA* e melhorar as taxas de morbimortalidade dos pacientes. Aqui, descrevemos um modelo de aprendizado de máquina para prever a ocorrência de HACs em cinco categorias distintas usando dados clínicos temporais. Usando nossa abordagem, descobrimos que pelo menos US$ 10 bilhões em custos hospitalares excessivos poderiam ser economizados somente nos *EUA*, com a instituição de medidas preventivas eficazes. Além disso, também identificamos vários recursos fundamentais que demonstram alto poder preditivo para HACs em diferentes períodos após a admissão do paciente. Os classificadores e os recursos analisados neste estudo mostram uma alta promessa de poder ser usado para previsão precisa de HACs em contextos clínicos, além de fornecer novas ideias sobre a contribuição de vários fatores clínicos para o risco de desenvolver HACs em função do sistema de saúde exposição.

Em conclusão, os HACs são um problema proeminente nos hospitais modernos e fornecem uma drenagem significativa no sistema de saúde. Classificadores de aprendizado de máquina como os desenvolvidos neste estudo, podem ser ferramentas valiosas para auxiliar os médicos na detecção, mitigação e finalmente na prevenção de HACs (WARNER *et al.*, 2016).

O artigo de Brossette e Hymel (2008) foi realizado uma análise da mineração de dados em medicina de laboratório e controle de infecções. Segundo os autores, o controle de infecções é a atividade de controle de qualidade relacionada principalmente à quantificação e prevenção de infecções hospitalares. Seu sucesso depende da identificação e correção oportunas de falhas no processo que aumentam os riscos de infecção. No entanto, é difícil para o controle de infecção identificar novas ameaças. Esses desafios podem ser mitigados por um sistema de mineração de dados projetado corretamente.

Para Brossette e Hymel (2008), a incapacidade dos profissionais de controle de infecção em identificar de maneira confiável as infecções hospitalares, muito menos padrões

entre elas, é uma limitação clara do sistema tradicional endossado pelo Centro de Controle de Doença (CDC).

Por esse motivo foi criado o Marcador de Infecção Hospitalar Eletrônico (NIM). O NIM supera as definições de casos clínicos do sistema nacional de vigilância de infecções hospitalares. O NIM é passível de ser computado, resolvendo assim uma grande limitação dos métodos manuais de busca de casos. Os modelos de dados baseados no NIM, podem descrever de maneira específica e confiável os padrões de infecções hospitalares, não apenas os resultados de laboratório, permitindo iniciativas de melhoria de processo mais específicas e objetivas (BROSSETTE; HYMEL, 2008).

Ainda conforme Brossette e Hymel (2008), o Sistema de Vigilância de Mineração de dados (DMSS) fornece uma ilustração prática da utilidade da mineração de dados na área da saúde. O acesso a dados eletrônicos adicionais pode ampliar os recursos e a utilidade de construção de modelos do DMSS.

Por exemplo, dados adicionais sobre a origem do paciente podem permitir que os modelos descrevam ou prevejam padrões significativos em casas de repouso, CEPs entre outros, dados adicionais, como procedimento cirúrgico, sala de cirurgia, tempo operatório, anestesia e classe da ferida, podem aumentar os padrões associados à cirurgia. Dados de uso de antimicrobianos ou hemograma completo podem aumentar a sensibilidade e a especificidade do secundário, mesmo que apenas para subconjuntos específicos de pacientes (BROSSETTE; HYMEL, 2008).

Conforme relatam Cánovas-Segura *et al.* (2016), as infecções causadas por bactérias e outros micro-organismos são um dos problemas de saúde mais relevantes no momento. As melhores soluções clínicas para esse problema são os antibióticos, drogas únicas devido à sua alta eficácia em termos de redução da mortalidade. Quando expostas a um ambiente antibiótico, as bactérias são capazes de desenvolver resistência rapidamente devido ao seu curto ciclo de crescimento e seus múltiplos mecanismos de adaptação.

O artigo de Cánovas-Segura *et al.* (2016) apresenta as necessidades de um Sistema de Apoio à Decisão Clínica (CDSS) orientado para ajudar os grupos de Programas de Administração Antimicrobiana (ASP) na tarefa de administração de antibióticos. Os autores relatam que uma combinação de regras de produção, ontologias, modelagem de fluxo de trabalho e técnicas de descoberta de subgrupos podem ser usadas para atender alguns requisitos, três requisitos mais importantes do cenário são: uma perspectiva multiusuário, para adaptar a resposta e as funcionalidades do CDSS a cada perfil de usuário; a capacidade de

trabalhar de maneira reativa, após a solicitação de um usuário e de maneira proativa, dando ao sistema a iniciativa de comunicar informações relevantes aos usuários; e a capacidade de incorporar conhecimento de diferentes tipos de fontes, como diretrizes clínicas, conhecimento especializado e até resultados de processos de mineração de dados.

Devido à sua flexibilidade foi escolhido o mecanismo de regras *Drools* e linguagem de programação Java. Os autores relatam que foi desenvolvida uma plataforma básica para administração de antibióticos. Esta plataforma está sendo testado e avaliado por médicos, onde os mesmos ajudarão a continuar aprimorando o conhecimento sobre o cenário (CÁNOVAS-SEGURA *et al.*, 2016).

Segundo Wiens, Guttag e Horvitz (2016), o aumento de Electronic Health Records ou em português Registros Eletrônicos de Saúde (EHRs) cria oportunidades para o uso de aprendizado de máquina para criar modelos que ajudam os profissionais de saúde a melhorar os resultados dos pacientes. Nos últimos anos, houve uma quantidade significativa de pesquisas dedicadas ao uso de dados clínicos para prever os resultados dos pacientes.

Como descrevem Wiens, Guttag e Horvitz (2016) em seu artigo, foram utilizados mais de 50.000 admissões de pacientes de um único hospital. Estes dados clínicos contêm informações sobre medicamentos, procedimentos, locais hospitalares, e equipe de saúde, resultados de laboratório, medições de sinais vitais, histórico do paciente e detalhes da admissão. A partir dessas informações buscou-se um mapeamento descrevendo um paciente, para uma estimativa da probabilidade do paciente de adquirir uma infecção.

3 PROJETO ISC

Este capítulo tem por objetivo apresentar a descrição da proposta do projeto de Infecção do Sítio Cirúrgico, assim como exibe o diagrama de caso de uso do sistema desenvolvido para este trabalho.

3.1 DESCRIÇÃO DA PROPOSTA

Inicialmente, vale destacar que este projeto foi desenvolvido no hospital da UNIMED de Criciúma, sul do estado de Santa Catarina, com aprovação do comitê de ética deste hospital, conforme Carta de Aceite no Apêndice A.

Desenvolvimento de uma ferramenta que possibilite a análise de risco de ISC de um paciente antes dele se submeter a um procedimento de saúde. Assim, tornando possível um reforço nas medidas preventivas, com o objetivo de preservar a vida e reduzir custos.

Esta análise se dará de forma automatizada e se baseará em dados existentes no prontuário/histórico do paciente e que foram analisadas e tratadas de forma estatística, gerando um modelo representação de comportamento, que poderia ser integrado a um sistema existente na unidade. Por meio desta ferramenta, é possível que um treinamento contínuo permita um aprimoramento contínuo com base na atualização de comportamentos.

Tal solução seria aplicável a toda unidade de saúde que conforme mencionado no capítulo 2, resultam em 80.000 mortes por ano e um custo estimado entre 4,5 e 5,7 bilhões de dólares. Com uma integração com um sistema existente, todos os envolvidos nos cuidados com o paciente poderiam adotar um comportamento diferenciado. Dependendo do procedimento e risco, o profissional poderia determinar que o procedimento não devia ser realizado na ausência de profilaxia, por exemplo. Talvez mudanças/evoluções nos protocolos existentes.

Além do uso prático no dia a dia de uma unidade, as análises observadas permitirão demonstrar a relevância em um tratamento mais adequado no registro de dados dos pacientes, visto que a identificação de comportamento se dá com base nos dados.

É relevante para o hospital determinar a causa das ISCs ou uma mudança na taxa, por exemplo, aumento. Dado que alguns comportamentos do próprio paciente podem ser responsáveis pelo desenvolvimento de uma complicação.

Atenção especial aos pós-procedimento, um dia a mais de internação, novos exames, retornar no consultório em três dias ao invés de sete como é o padrão.

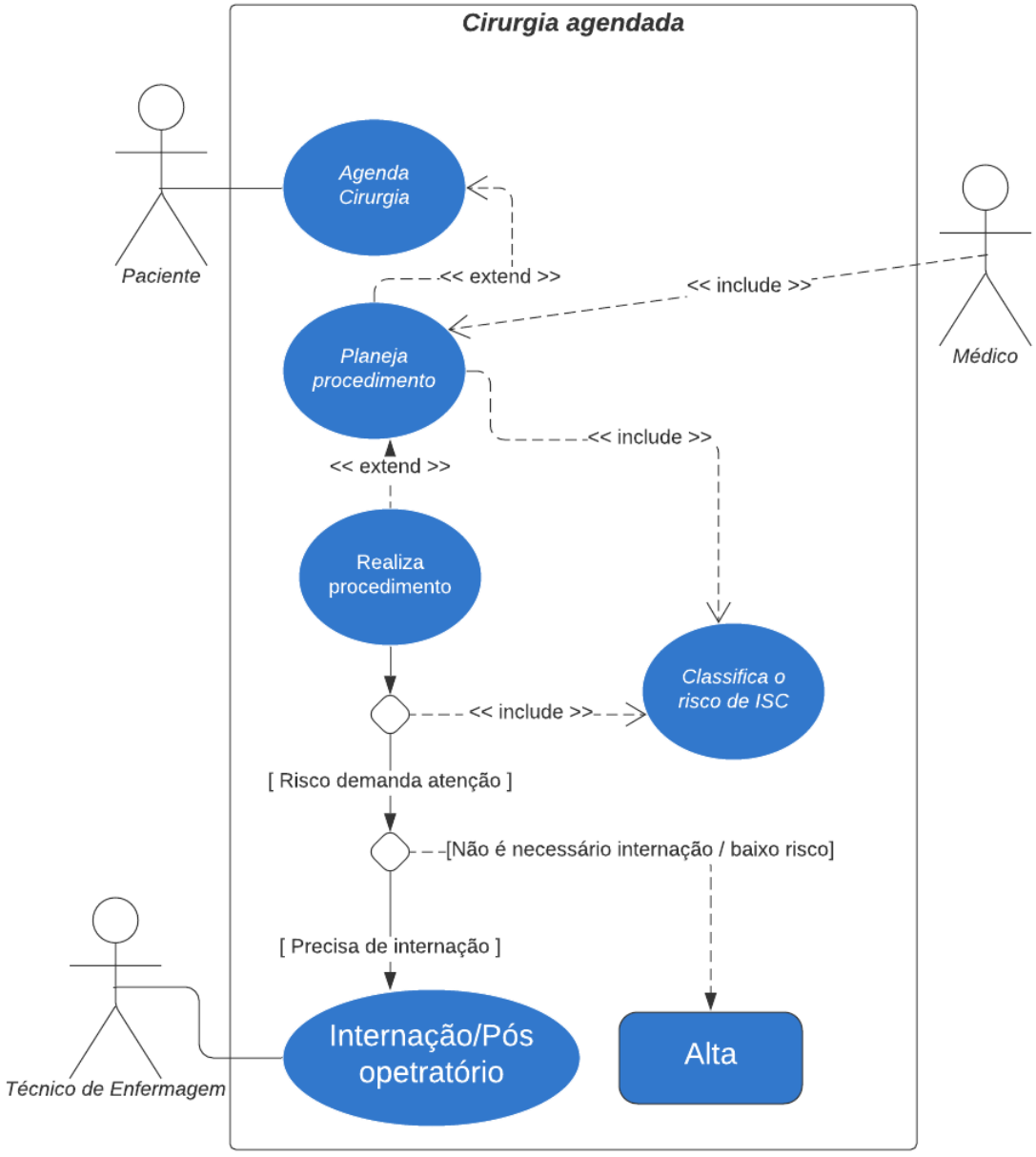
Na prática, o Sistema proposto faz um cruzamento de informações através de técnicas de Mineração de dados e a utilização de IA na qual aponta os possíveis riscos de infecção pós-cirúrgica. De forma mais clara e detalhada, o paciente necessita de cirurgia, na qual é agendada e o mesmo passa por uma consulta pré-anestésica onde o médico anestesista avalia todo o histórico, comorbidades, tipo de procedimento a ser realizado e anestesia. Logo após essa etapa será agendado o procedimento cirúrgico, e após a cirurgia deve-se realizar uma nova avaliação de acordo com o procedimento, analisando o nível de sucesso da cirurgia ou se houve algum tipo de complicação, em caso de alguma complicação havendo possíveis riscos, medidas devem ser tomadas, como estender o período de acompanhamento de internação. Do contrário, a cirurgia sendo um sucesso o paciente pode receber alta. Dentre as formas de avaliar no pós-operatório, além da avaliação médica o técnico em Enfermagem acompanha diariamente o paciente, monitorando sinais vitais e ministrando medicamentos.

Nessa rotina o Técnico em Enfermagem antes de se deslocar ao quarto do paciente, pode checar os dados conferindo o diagnóstico de risco de infecção, além disso, existem computadores com acesso ao sistema TASY que é um sistema de Gestão Hospitalar, onde todas as informações ficam disponíveis para consulta. No procedimento de alta Hospitalar, se incluem alguns regramentos de acordo com o risco de Infecção do paciente, onde os retornos para consulta pós-cirúrgicas são padronizadas em sete dias, para os pacientes com riscos de Infecção sugere-se retornar em três dias.

O sistema de classificação de risco de ISC ficará integrado junto ao sistema TASY e será utilizado nos processos de cirurgias agendadas e cirurgias de emergência. Os principais atores no processo são o paciente, médico e técnico em enfermagem. Abaixo, o caso de uso de cirurgias agendadas. Lembrando que os procedimentos já existentes não serão alterados com as informações antes da cirurgia, mantém o cumprimento de todos os protocolos como: técnica asséptica, uso do antibiótico profilático e o que este projeto sugere é de que o sistema *TASY* alerte a equipe (médicos, enfermeiros e técnicos) para que se tenha mais atenção aos sinais pós operatório: por exemplo um paciente retirou o apêndice – apendicectomia relata dor e o sistema apontou grande chance de infecção, o sistema alertará para que os cuidados e atenção sejam redobrados, não autorizando alta para o paciente como é de costume e reavaliando com novos exames.

Para melhor compreensão do sistema desenvolvido para este projeto a Figura 29 expõe o diagrama de caso de uso, tendo como atores: Paciente, Médico (a) e Técnico (a) de enfermagem/Enfermeiro (a), bem como as ações são: Cirurgia agendada, Cirurgia de emergência e Estadia no hospital.

Figura 29 - Diagrama de Caso de Uso



Fonte: Elaborado pelo AUTOR (2021).

4 PROCEDIMENTOS DE DESENVOLVIMENTO

Neste capítulo são abordadas as ferramentas utilizadas, bem como, suas respectivas etapas no desenvolvimento do sistema. Na segunda seção são apresentadas as análises realizadas em cada etapa do projeto.

4.1 FERRAMENTAS UTILIZADAS NO DESENVOLVIMENTO

Para o projeto foi utilizada a linguagem de programação *Python* na versão 3.7.3 em um ambiente de desenvolvimento Ubuntu. Antes de iniciar a análise foi necessário ler o arquivo com os dados coletados no formato xlsx. Foi utilizada a biblioteca Pandas, onde teve início a etapa de limpeza dos dados. Foram removidas linhas em branco, efetuadas algumas tratativas nas *features* numéricas. Nesta etapa foram feitas algumas visualizações exploratórias com os pandas em formato de tabelas, mas com a necessidade de algumas visualizações mais avançadas foi empregado o *matplotlib* em conjunto com o *seaborn*.

As visualizações utilizadas foram métodos da estatística descritiva, como média, mínimos, máximos e desvio padrão de algumas features. Além de *boxplots* para identificar *outliers* causados por algum erro no preenchimento por parte do médico ou até mesmo na coleta dos dados.

Algumas *features* como comorbidade, cirurgia, tipo de antibiótico, estavam disponíveis no prontuário do paciente de forma textual, podendo existir vários valores em uma única célula, onde foi necessário separá-las em variáveis binárias e/ou categóricas. Para isso, foram utilizadas funções da própria linguagem *Python* em conjunto com a biblioteca *unicodedata* para remover os acentos das palavras e unificar possíveis variações causadas por erros de preenchimento.

O Quadro 1 apresenta as etapas e ferramentas utilizadas no desenvolvimento.

Quadro 1 – Etapas e ferramentas utilizadas no desenvolvimento

Etapa	**Ferramentas**
Coleta dos dados e análise do problema	Sistema *Tasy*
Limpeza dos dados e análise exploratória	Python, Pandas, *Matplotlib, Seaborn*, Unicodedata
Feature engineering	Python, Pandas, *Numpy*

Treinamento e testes	● *Scikit-learn:*
	  ○ *Random Forest Classifier;*
	  ○ *Extra Trees Classifier;*
	  ○ *Gradient Boosting Classifier;*
	  ○ *Logistic Regression;*
	  ○ *Train_test_split;*
	  ○ *Grid Search CV;*
	  ○ *Stratified KFold.*
	● *Xgboost.*

Análise dos resultados	● *Scipy stats;* ● *confusion_matrix;* ● *accuracy_score;* ● *average_precision_score;* ● *recall_score;* ● *precision_recall_fscore_support;* ● *Heatmaps* para correlação.

Fonte: Elaborado pelo AUTOR (2021).

4.2 ANÁLISES REALIZADAS

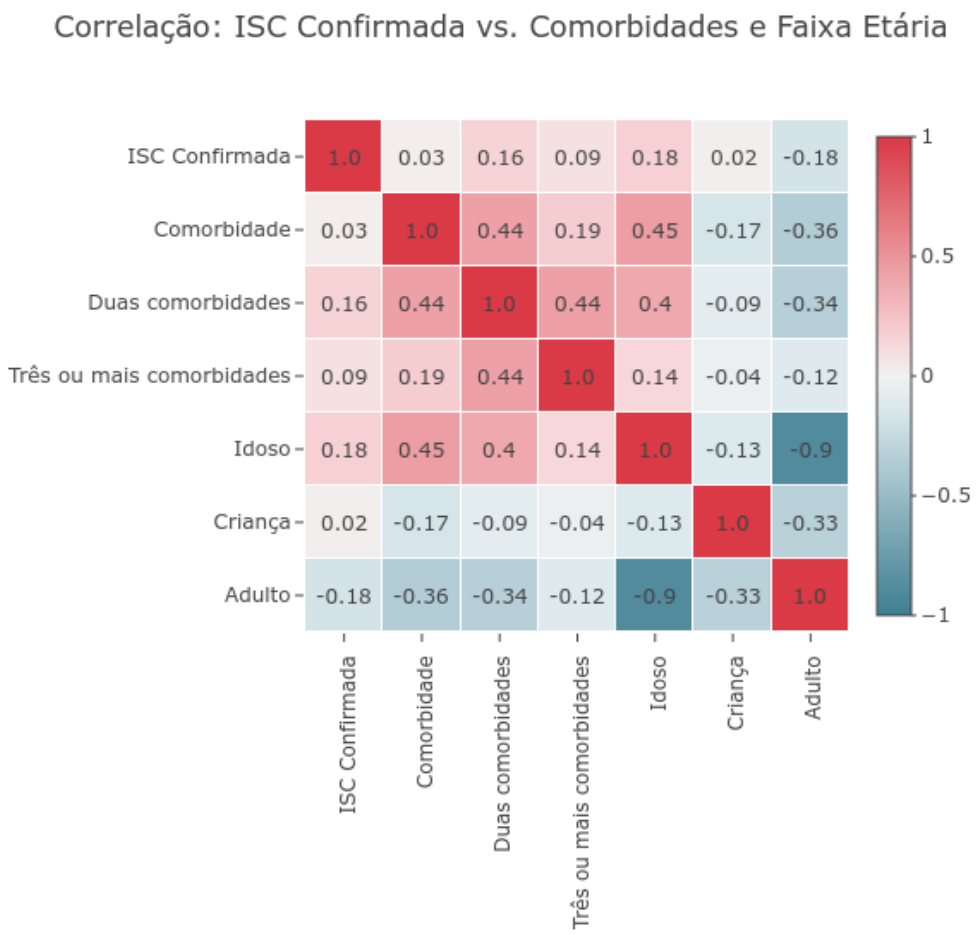
Na sequência, foi realizada a *feature engineering*, com base nas análises realizadas, onde foram criadas faixas de valores para algumas *features*, como o tempo de cirurgia, onde ao invés de usar um valor contínuo, foram usadas as binárias: até 30 minutos, até 60 minutos, até 90 minutos, etc. Assim, foi evitada a dispersão que poderia ocorrer em razão do valor numérico.

Outra *feature* onde foi necessário tratar a dispersão foi os casos das comorbidades. Alguns tipos possuíam poucas amostras, conversando com os especialistas, foi possível agrupar as comorbidades e reduzir o risco de algumas terem um peso alto para classificação de uma infecção ou não. Por exemplo, existem várias cardiopatias e doenças respiratórias.

Foram plotadas as correlações das *features* para determinar a relevância e as interações entre elas. Este processo foi realizado inicialmente antes de realizar a feature engineering e também após a criação de algumas features. Assim, foi possível ver, por exemplo, como o agrupamento das comorbidades afetou as correlações e interações, reduzindo o peso de alguma comorbidade ou procedimento específico.

Na Figura 30, estão demonstradas as interações entre o número de comorbidades e faixa etária, na coluna ISC Confirmada. É possível ver que até uma comorbidade, existe uma pequena correlação positiva com ISC, sendo que duas, resulta em uma correlação positiva quase cinco vezes maior.

Figura 30- Correlação: ISC Confirmada vs. Comorbidades e faixa etária.

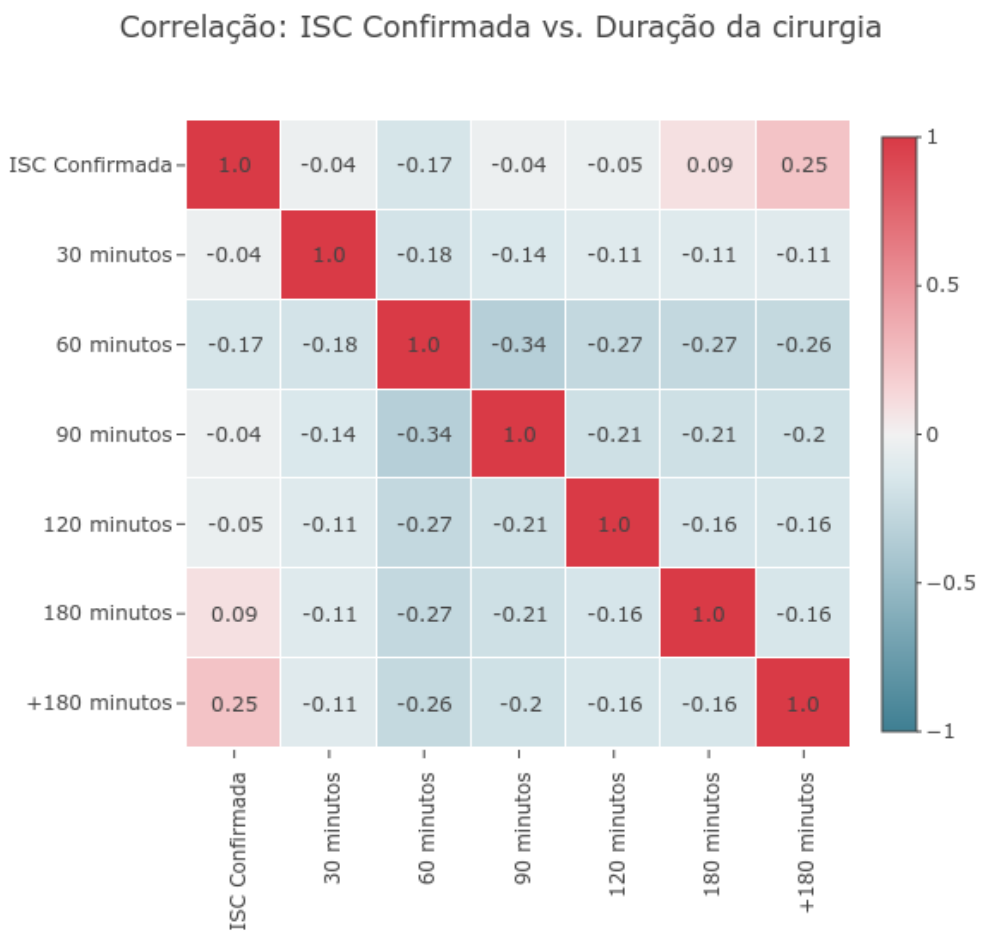


Fonte: Elaborado pelo AUTOR (2021).

Pacientes idosos também tiveram uma alta relação com o desenvolvimento de uma ISC, onde adultos apresentaram uma correlação negativa. A relação entre pacientes idosos e a existência de comorbidades também foi uma das mais altas, cerca de 0,45.

Cirurgias de até 120 minutos apresentaram uma correlação negativa com a ISC. O tempo de cirurgia também deve estar relacionado com a sua complexidade. Cirurgias com mais de 120 minutos até 180, tiveram uma correção de 0,09 com a confirmação de uma ISC, enquanto que nas cirurgias com mais de 180 minutos, obteve-se o maior valor. A Figura 31 apresenta a correlação ISC confirmada vs. duração da cirurgia.

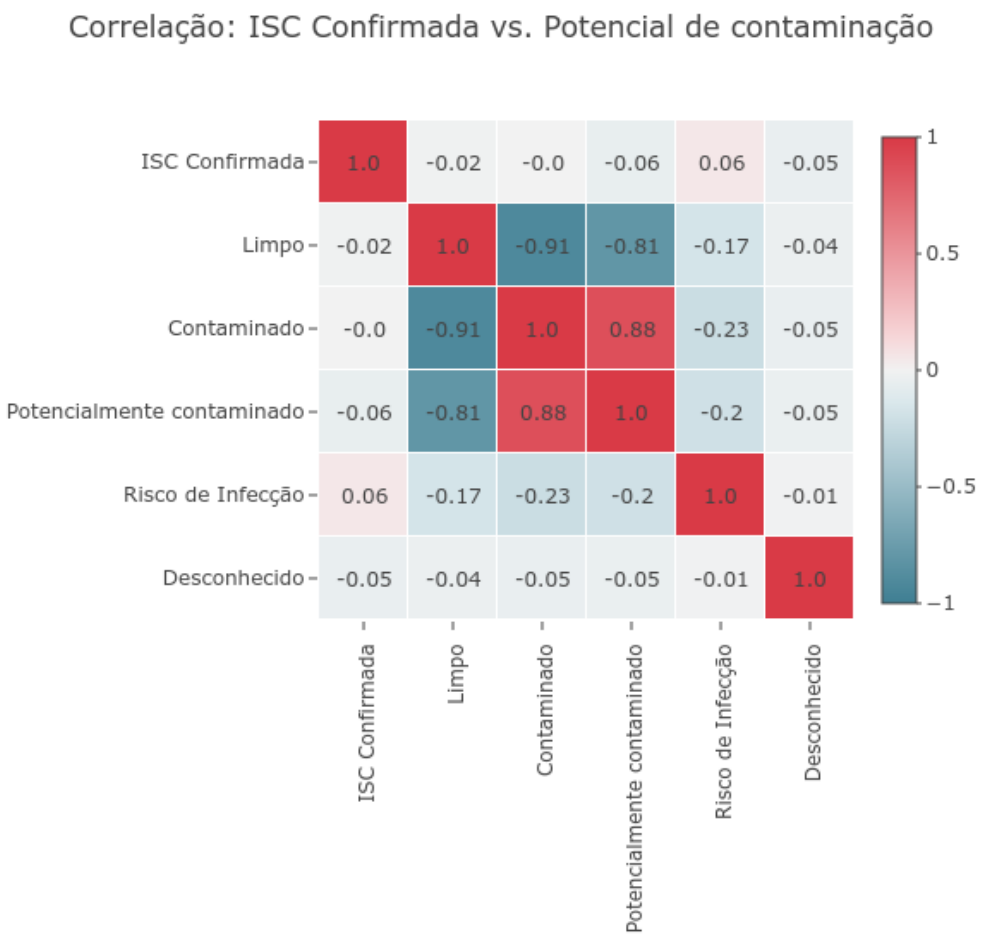
Figura 31 - Correlação: ISC Confirmada vs. Duração da cirurgia.



Fonte: Elaborado pelo AUTOR (2021).

O potencial de contaminação nos cenários limpo, contaminado, potencialmente contaminado e desconhecido, demonstraram uma correlação nula ou negativa, como pode ser observado na Figura 32. Sendo a única interação positiva quando o procedimento foi descrito com risco de infecção.

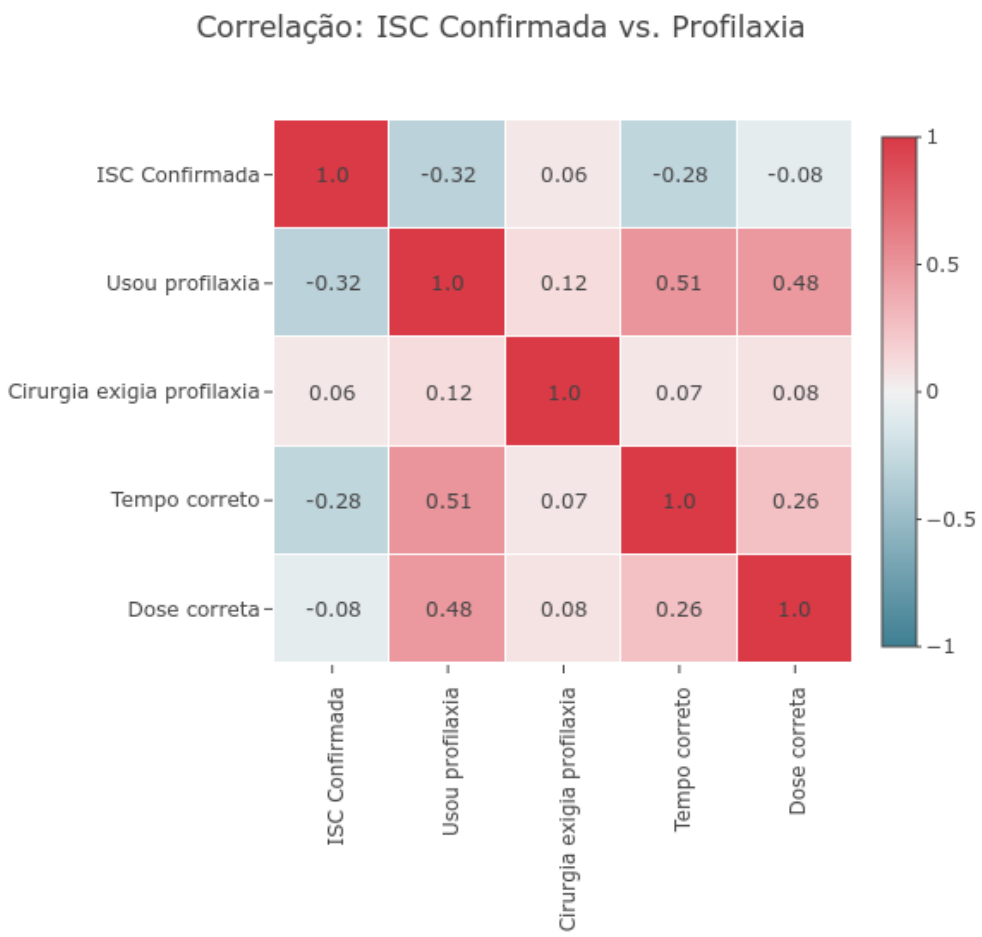
Figura 32 - Correlação: ISC Confirmada vs. Potencial de contaminação.



Fonte: Elaborado pelo AUTOR (2021).

Para avaliar a relação entre ISC, necessidade e uso de profilaxia, foram plotadas as interações entre estes atributos. Na Figura 33, é demonstrada uma correlação positiva com ISC enquanto que o uso de profilaxia possui uma interação negativa e significativa. Demonstra que o uso de profilaxia pode ter ajudado na redução do número de ISCs. As variáveis tempo correto e dose correta reforçam essa provável redução.

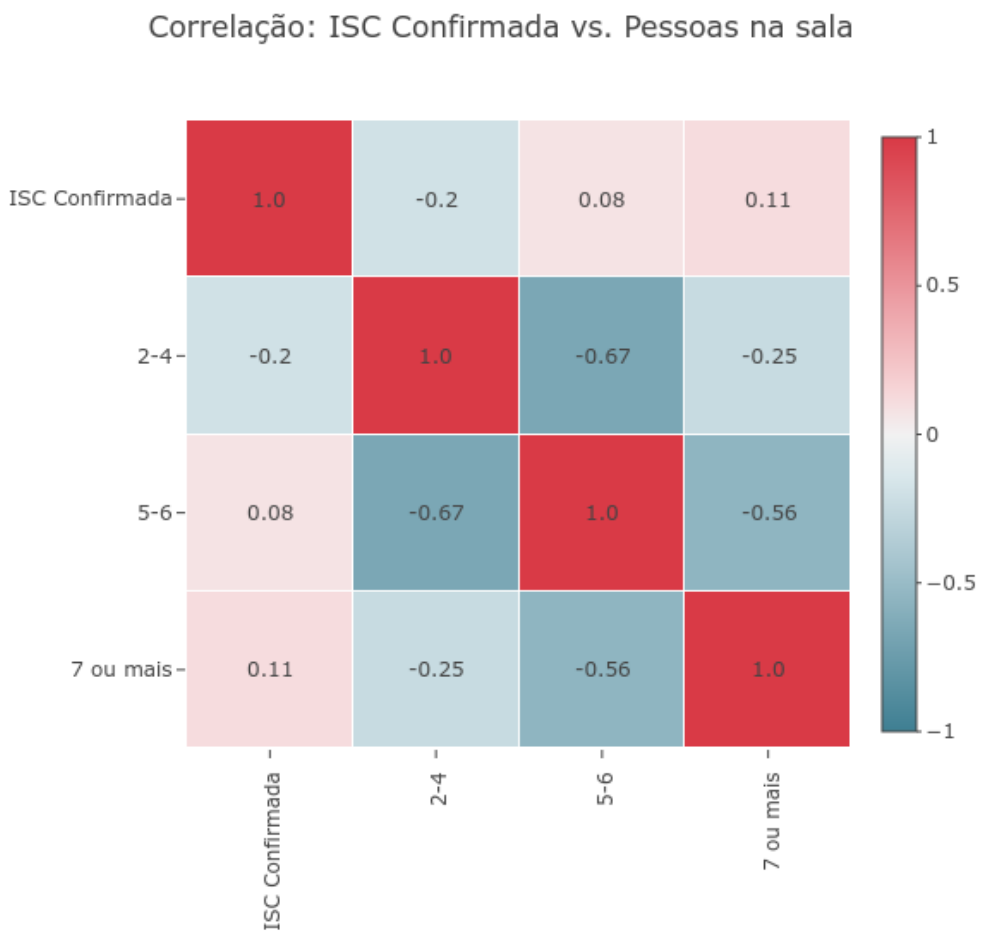
Figura 33 - Correlação: ISC Confirmada vs. Profilaxia



Fonte: Elaborado pelo AUTOR (2021).

Na Figura 34, estão demonstradas as interações entre o desenvolvimento de uma ISC e o total de pessoas na sala de cirurgia. Quando existem na sala de duas a quatro pessoas, existe uma interação negativa. De cinco a seis, existe uma interação positiva quatro vezes maior, chegando a ser quase seis vezes maior quando o total de pessoas na sala é de sete ou mais pessoas.

Figura 34 - Correlação: ISC Confirmada vs. Pessoas na sala.

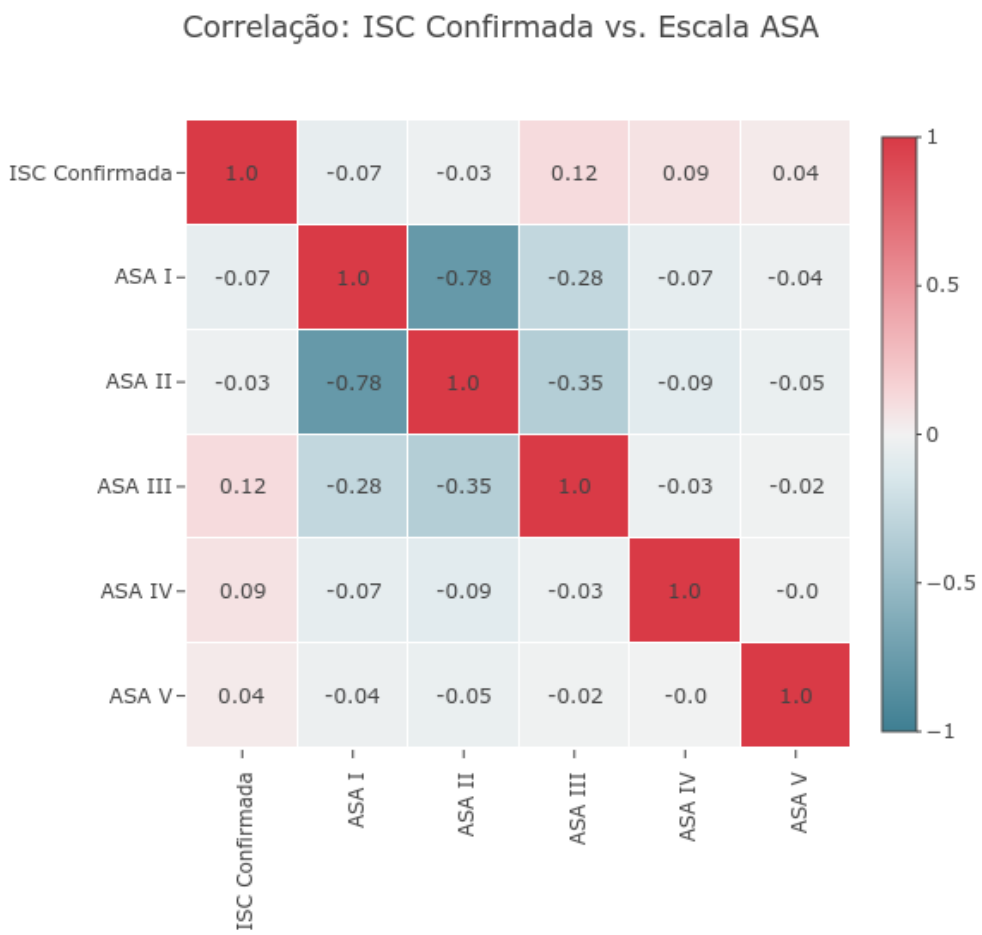


Fonte: Elaborado pelo AUTOR (2021).

A escala ASA classifica o risco do procedimento para um paciente de acordo com o seu estado de saúde.

Pensando nisso, foi avaliada a classificação ASA e sua interação com o desenvolvimento de uma ISC, Figura 35. Na escala I e II, houve uma correlação negativa. Enquanto que na III, IV, e V, as de risco mais elevado, observa-se as interações positivas mais elevadas.

Figura 35 - Correlação: ISC Confirmada vs. Escala ASA.



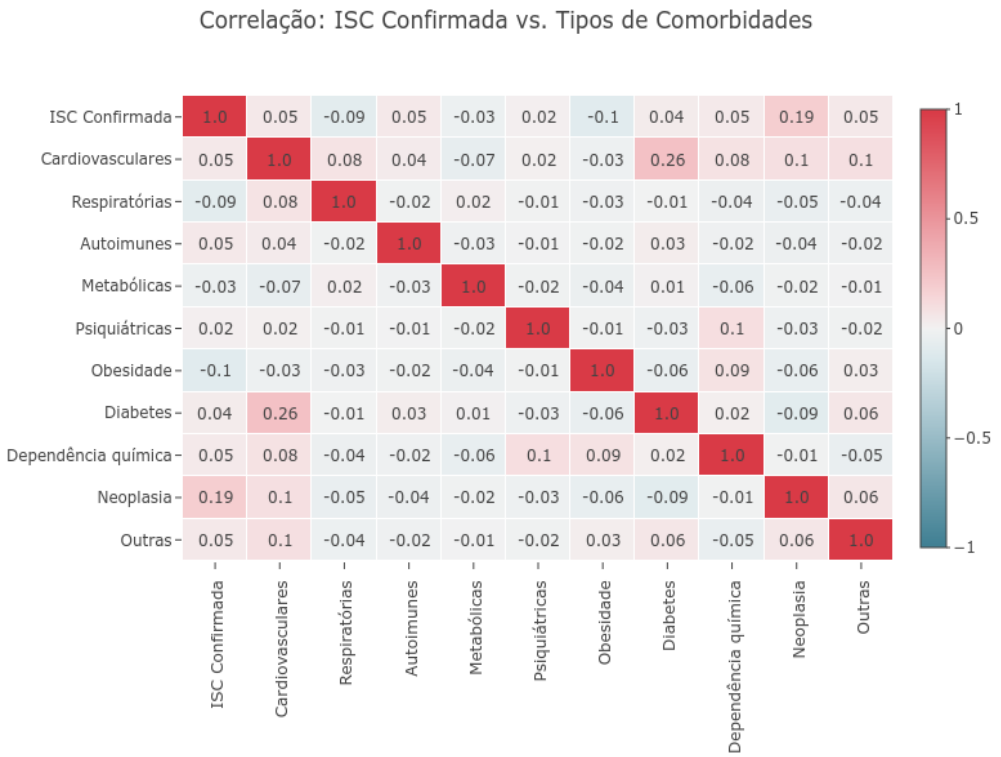
Fonte: Elaborado pelo AUTOR (2021).

A correlação positiva com a escala V, que demonstra urgência na operação do paciente, pode ser justificada por não existir uma profilaxia antes, dada a necessidade da rápida operação.

Em relação aos tipos de comorbidade, a interação positiva e mais forte foi observada com neoplasia, um tipo de câncer, seguida por cardiovasculares, autoimunes, dependência química, diabetes. O agrupamento em “outras”, são algumas comorbidades onde não existia uma amostragem significativa das mesmas. Por isso, não é possível fazer qualquer afirmação sobre as mesmas.

O agrupamento das comorbidades também contribuiu para que o modelo uma *feature* que regularizasse o fator preditivo de infecções hospitalares. A Figura 36 expõe a correlação ISC confirmada vs. tipos de comorbidades.

Figura 36 - Correlação: ISC Confirmada vs. Tipos de Comorbidades.



Fonte: Elaborado pelo AUTOR (2021).

Um procedimento X poderia pela amostragem dos dados contêm uma taxa de infecção de 50%, quando no mundo real, ele seria de 2%.

Também se observou uma correlação negativa com comorbidades respiratórias e obesidade. No caso de obesidade, a interação faz sentido se considerarmos a baixa taxa de infecção em procedimentos relacionados a esta comorbidade, cerca de 0,3% (Apêndice B).

Após avaliação das interações entre as *features*, foi iniciado o treinamento do modelo de classificação de ISC. Primeiramente foram criados métodos que facilitassem a otimização de parâmetros de acordo com o algoritmo utilizando o *Grid Search CV*. Além disso, alguns parâmetros como o percentual de dados utilizado para treinamento e teste e o parâmetro de *seed* utilizado nas operações de *split*, para que fosse possível reproduzir os resultados, independente do ambiente onde o software fosse executado.

Ao final de cada treinamento, era gerado um arquivo com as métricas calculadas para cada algoritmo, além do melhor resultado e quais parâmetros foram utilizados.

Com estes arquivos de resultado, foi possível ao final de todos os treinamentos e validação, carregar os resultados e comparar os algoritmos e o quanto eles foram eficientes,

determinando qual o melhor, que manteve um balanço entre os falsos positivos e falsos negativos.

As métricas utilizadas durante o treinamento tinham como objetivo obter o melhor classificador considerando *tradeoff* entre a *precision* e *recall*. Como mencionado anteriormente, o objetivo de um classificador é identificar o máximo possível de casos da classe positiva (infecção = 1). Porém, gerar falsos positivos, pois em um cenário real, os esforços para que seja tomada uma ação com base nos algoritmos seria limitado (ex: dedicar um cuidado diferenciado com alguns pacientes cujo algoritmo informou que ele iria desenvolver uma infecção). Um classificador que chutasse sempre que todo paciente iria desenvolver uma infecção, acertaria 100% das vezes, mas iria gerar um desperdício de recursos ou fazer com que recursos fossem alocados de forma equivocada.

5 ANÁLISE E DISCUSSÃO DOS RESULTADOS OBTIDOS

Este capítulo busca descrever, na primeira seção, os resultados da pesquisa com os referidos dados apurados. Na segunda seção, têm-se a análise do problema com a apresentação dos dados da unidade de saúde objeto de pesquisa deste trabalho. Na terceira seção é abordado as contribuições do trabalho em comparação com o projeto do Robô Laura e *Watson Health*.

5.1 RESULTADOS

Os algoritmos foram ranqueados com base na relação entre a *precision* e o *recall*, para priorizar a detecção de casos de ISC e minimizar o total de falsos positivos, já que um falso positivo no mundo real poderia significar um desperdício de recursos com um paciente que não teria uma infecção.

Com este critério, na Figura 37 pode ser visto que o *Gradient Boosting Classifier* e *Logistic Regression* obtiveram os melhores resultados, ficando empatados com um *Recall* de 79% e uma acurácia de quase 74%.

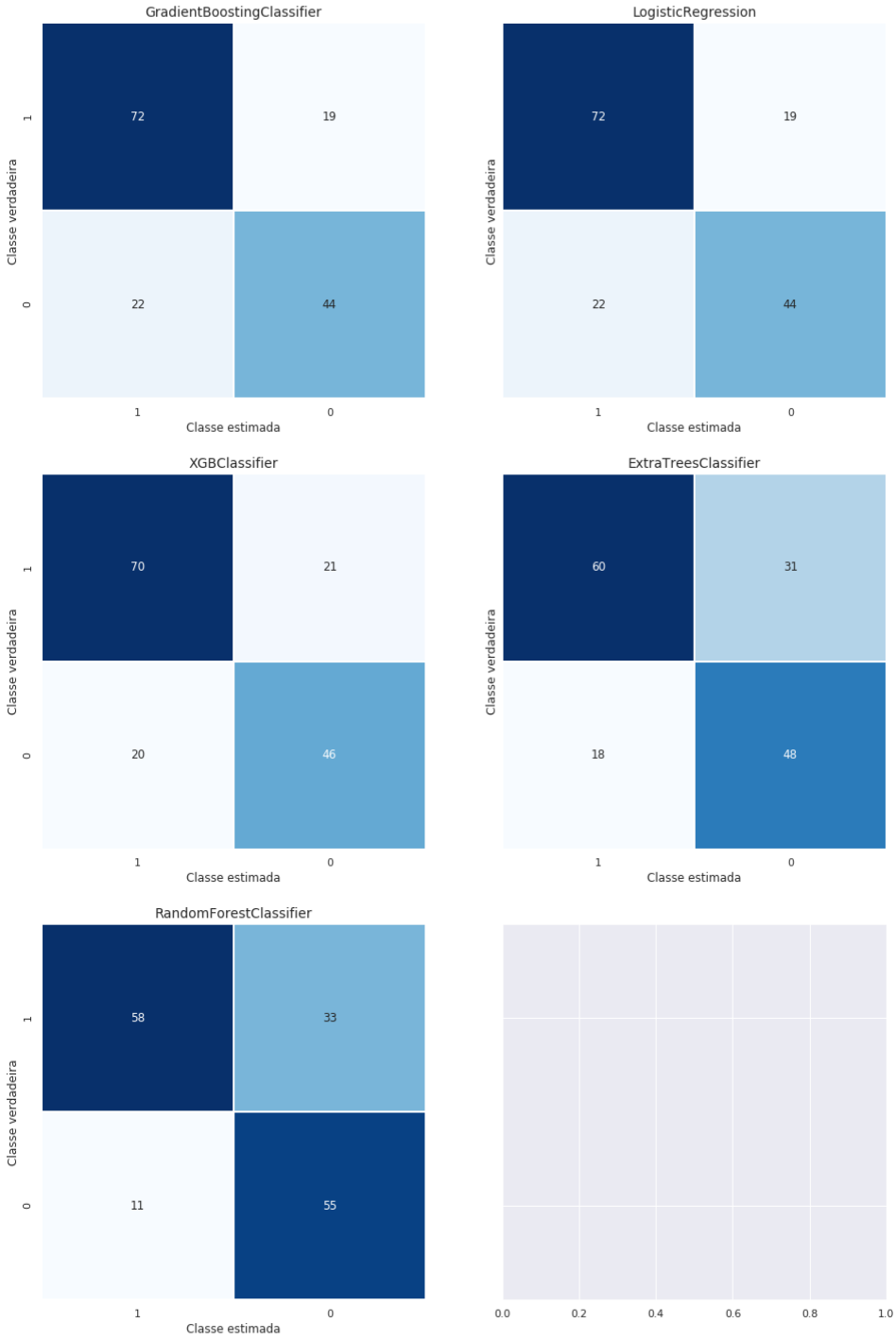
Figura 37 - Resultados do Gradiente Boosting Classifier e Logistic Regression

Modelo	Recall ▼	Acurácia	AUC	Tempo de treinamento (s)
Gradient Boosting Classifier	0.791	0.739	0.729	1057
Logistic Regression	0.791	0.739	0.729	23
XGB Classifier	0.769	0.739	0.733	8175
Extra Trees Classifier	0.659	0.688	0.693	22693
Random Forest Classifier	0.637	0.720	0.735	13720

Fonte: Elaborado pelo AUTOR (2021).

A Figura 38 apresenta a matriz de confusão calculada com os dados de teste para cada um dos algoritmos avaliados, onde pode ser visto que com o *Gradient Boosting* e *Logistic Regression*, 72 das 91 ISCs existentes. Outro detalhe é que ocorreram 22 falsos positivos de 66 não infecções, previstas como ISCs.

Figura 38 - Matriz de confusão calculada

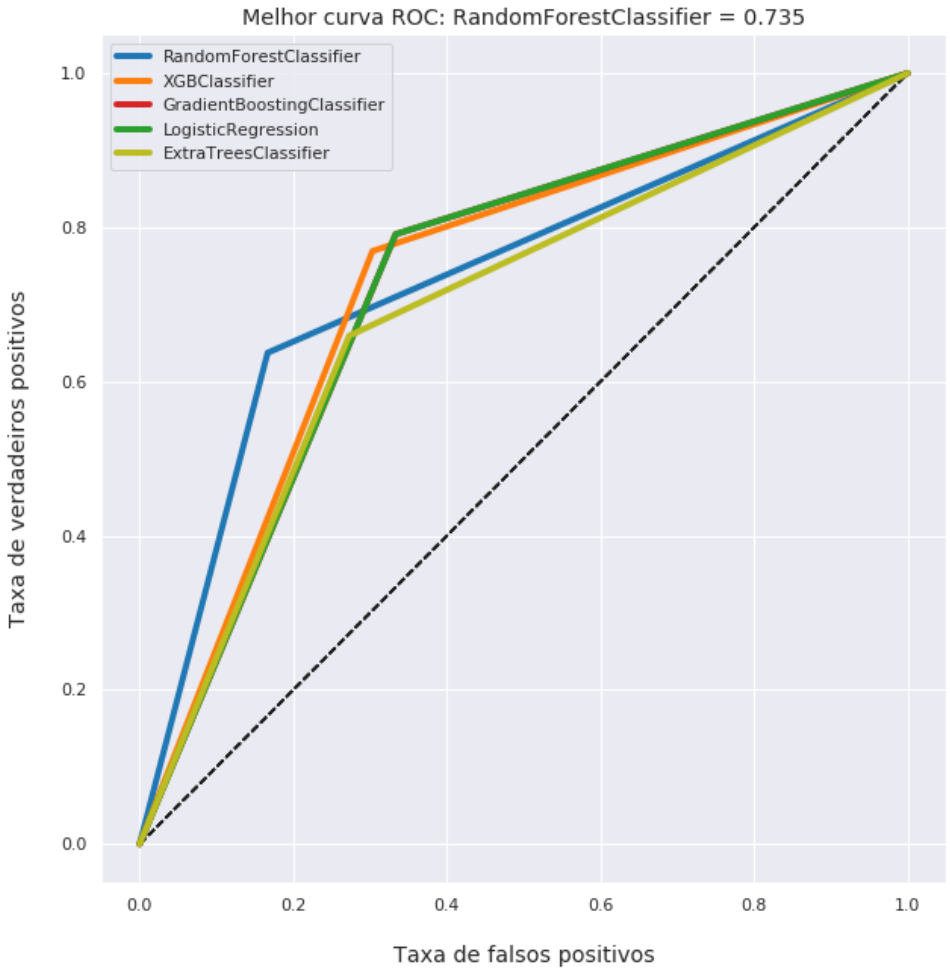


Fonte: Elaborado pelo AUTOR (2021).

Na Figura 39 foi calculada também a curva ROC para os algoritmos avaliados. Ela demonstra a relação entre a detecção de ISCs e falsos positivos, que na vida real podem significar recursos alocados em cuidados especiais para pacientes que não irão desenvolver

ISC. Pode-se notar que o ROC do *Gradient Boosting* é igual ao do *Logistic Regression*, as duas linhas estão sobrepostas.

Figura 39 - Resultado do ROC do gradiente boosting e logistic regression



Fonte: Elaborado pelo AUTOR (2021).

A melhor Curva observada foi do *Random Forest Classifier*, porém, o mesmo teve um baixo *recall* e uma diferença de menos de 1%. Isso mantém o *Gradient Boosting Classifier* e *Logistic Regression* como as melhores soluções obtidas dados seu *recall* de quase 80%.

Em termos de performance e tempo de treinamento, o *Gradient Boosting* obteve o mesmo resultado que o *Logistic Regression*, porém, foi mais lento. *Logistic Regression* teve

um tempo de treinamento de 23 segundos para validar 3 mil combinações, enquanto que o *Gradient Boosting* levou 18,5 minutos para testar 105 mil combinações.

Exemplo: Os dois melhores classificadores previram com sucesso cerca de 79% das infecções hospitalares, com uma acurácia de 73%.

De acordo com Brasil (2017), nota-se uma crescente no número de intervenções cirúrgicas na assistência à saúde, reflexo do aumento de expectativa de vida e violência, que tem no ranking de cirurgias as doenças cardiovasculares, neoplasias e traumas.

Dados apontam para 187 a 281 milhões de cirurgias anualmente, em média uma cirurgia para cada 25 habitantes, dentro desse número considerável, vê-se as complicações de procedimentos cirúrgicos que variam de 3% a 16% e os óbitos que alcançam a margem de 5% a 10%, sendo um dado alto e assim preocupando a saúde pública (BRASIL, 2017).

Brasil (2017) ainda completa que, nos EUA estima-se de 150 a 300 mil infecções de Sítio Cirúrgico, responsáveis por mais de oito mil óbitos por ano. A ISC tem sido considerada como a IRAS mais comum e de maior custo, sendo que até 60% delas são passíveis de prevenção se adotar medidas sugeridas pelo Guidelines. Para Boxwala *et al.* (2001) e Santiago (2008), os Guidelines se referem a condutas e procedimentos desenvolvidos sistematicamente para auxiliar ao médico em decisões relativas à melhor e apropriada conduta em determinadas situações clínicas, e segundo a Associação Médica Brasileira, é Diretriz. No Brasil infelizmente não se tem dados sistematizados e aprofundados, mas se coloca em terceiro lugar no conjunto IRAS, com 14% a 16% dos pacientes hospitalizados.

Infecções do Sítio Cirúrgico são consideradas eventos adversos, complicações comuns no ato cirúrgico, que ocorrem no pós-operatório e que afetam em média de 3% a 20% dos procedimentos, impactando na morbidade e mortalidade dos pacientes, causando dano físico, social e/ou psicológico, sendo uma ameaça à segurança do paciente. Além de tais prejuízos físicos, psicológicos e financeiros ao paciente, as ISC costumam prolongar a estadia numa média de 7 a 11 dias, aumentando também a chance de readmissão hospitalar, cirurgias adicionais, e por consequência, elevando os gastos com o tratamento nos quais podem chegar a US$ 1,6 bilhões anuais (BRASIL, 2017).

Diante de dados impactantes e analisando a possibilidade de evitar tais danos, torna-se imprescindível implementar medidas urgentes de prevenção das ISC por meios diversos como protocolos, guias, manuais com base científica, medidas e listas de verificações que sejam relevantes para a redução dessas taxas de ISC (BRASIL, 2017).

O presente estudo vem mostrar os dados e medidas no intuito de contribuir para a redução da incidência das IRAS nos serviços de saúde.

Na Tabela 1 apresentam-se os dados de Infecções no centro cirúrgico. No entanto, além desse, existem outras infecções que se encontram circulantes dentro de um hospital.

Tabela 1 - Dados de infecções no centro cirúrgico

DESPESAS	jan/19	fev/19	mar/19	abr/19	mai/19	jun/19	jul/19	ago/19	set/19	out/19	nov/19	dez/19	TOTAL
Materiais	R$ 12.403,10	R$ 2.923,66	R$ 10.964,25	R$ 18.521,29	R$ 49.398,38	R$ 15.431,54	R$ 1.913,79	R$ 32.997,51	R$ 58.512,77	R$ 1.845,36	R$ 2.704,18	R$ 253,28	R$ 207.869,11
Antibiótico	R$ 4.647,23	R$ 9.306,39	R$ 12.679,26	R$ 47.306,58	R$ 44.826,05	R$ 9.350,99	R$ 7.325,92	R$ 30.576,94	R$ 62.376,69	R$ 13.251,05	R$ 9.024,48	R$ 1.493,87	R$ 252.165,45
Outros Medicamentos	R$ 25.978,13	R$ 3.229,68	R$ 9.128,85	R$ 29.085,97	R$ 28.180,17	R$ 5.207,35	R$ 3.481,80	R$ 27.372,78	R$ 50.070,96	R$ 9.871,85	R$ 5.385,42	R$ 205,51	R$ 197.198,47
Exames laboratoriais	R$ 1.682,39	R$ 804,83	R$ 988,09	R$ 2.910,56	R$ 3.200,30	R$ 1.381,07	R$ 673,29	R$ 5.089,86	R$ 6.155,40	R$ 426,59	R$ 561,22	R$ 61,73	R$ 23.935,33
Exame de imagem/ECG	R$ 1.039,02	R$ 1.221,22	R$ 2.617,03	R$ 4.516,68	R$ 4.244,34	R$ 190,00	R$ 538,34	R$ 1.731,80	R$ 7.201,37	R$ 1.742,57	R$ 1.275,62	R$ 0,00	R$ 26.317,99
Diárias/Taxas de sala/Visitas/Atendimento médico	R$ 45.442,49	R$ 22.142,99	R$ 36.389,46	R$ 69.825,57	R$ 131.695,14	R$ 8.215,74	R$ 20.620,93	R$ 41.922,79	R$ 191.674,51	R$ 15.509,11	R$ 21.648,15	R$ 1.998,61	R$ 607.085,49
Outras taxas/Gases medicinais/Nutrição/Fisioterapia/Tratamento de pele	R$ 3.382,61	R$ 2.299,44	R$ 3.503,87	R$ 13.060,98	R$ 22.177,74	R$ 1.450,39	R$ 686,55	R$ 2.863,83	R$ 67.409,65	R$ 1.820,07	R$ 5.611,58	R$ 494,62	R$ 124.761,33
DESPESA TOTAL/MÊS	**R$ 94.574,97**	**R$ 41.928,21**	**R$ 76.270,81**	**R$ 185.227,63**	**R$ 283.722,12**	**R$ 41.227,08**	**R$ 35.240,62**	**R$ 142.555,51**	**R$ 443.401,35**	**R$ 44.466,60**	**R$ 46.210,65**	**R$ 4.507,62**	**R$ 1.439.333,17**

Fonte: Elaborado pelo AUTOR (2021).

Foi possível identificar aproximadamente 80% dos casos de infecção do sitio cirúrgico. Os dois melhores classificadores previram com sucesso cerca de 79% das infecções hospitalares, com uma acurácia de 73%.

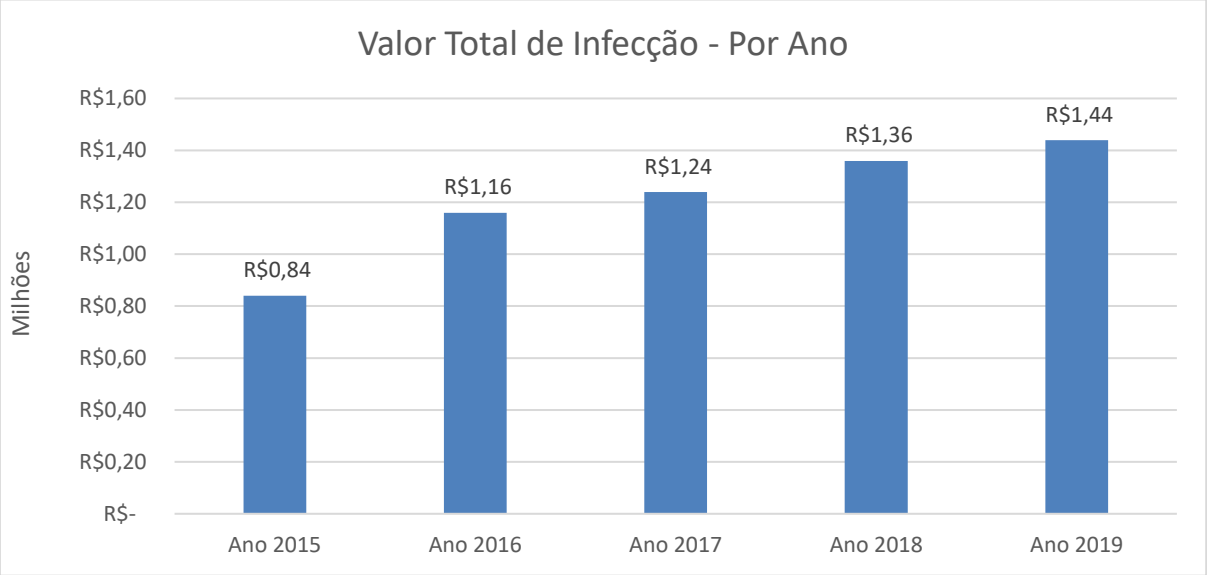
Na Tabela 2 ficam claro todas as formas possíveis de despesas divididos em períodos mensais, o que já nos mostra em números um valor de custo considerável para infecções e, por fim a tabela ainda apresenta um valor total anual que um hospital gera para tratar infecções, sendo que esse valor poderia ser muito melhor utilizado para fins de investimentos. No entanto este trabalho vem apresentar um meio de mudar essa realidade através do uso de inteligência artificial.

Tabela 2 - Levantamento de gastos com infecções nos últimos cinco anos

Ano	**Nº de infecção**	**Taxa de infecções**	**Total de procedimentos**	**Custo total Ano**
Ano 2015	21	0.1973	10.642	R$ 839.611,08
Ano 2016	29	0.2639	10.987	R$ 1.159.462,92
Ano 2017	31	0.2666	11.625	R$ 1.239.425,88
Ano 2018	34	0.2908	11.691	R$ 1.359.370,32
Ano 2019	36	0.2835	12.697	R$ 1.439.333,28

Fonte: Elaborado pelo AUTOR (2021).

Figura 40 - Valor total de infecções por ano



Fonte: Elaborado pelo AUTOR (2021).

No período de cinco anos:

- 151 infecções;
- R$ 6.037.203,48;
- Custo médio de uma infecção: R$ 39.981,48.

Com base nos custos para tratamentos de infecções nos anos 2015, 2016, 2017, 2018 e 2019 com o sistema proposto prevendo 79% das infecções seria possível reduzir os gastos em até R$4.829.736,00. Conforme informações e relatos dos profissionais atuantes no hospital onde o estudo foi realizado, os procedimentos elencados para prevenção das infecções já existem. Sendo assim, não haveria necessidade de investimentos e mudanças significativas de alto custo para prevenção das infecções, mas sim, a ferramenta proposta poderia servir para otimizar onde e quando os esforços seriam aplicados. Eventualmente, gastos adicionais poderiam ocorrer, como no caso de internação do paciente ser prolongada ou de algum medicamento adicional, mas é sabido que financeiramente os custos de uma infecção são significativamente maiores, além é claro do risco de morte.

5.2 ANÁLISE DO PROBLEMA

Através de acessos de dados na unidade de saúde onde o estudo se realizou, houve um levantamento de informações para que assim pudesse utilizar a Inteligência Artificial no Sistema proposto para alertar sobre possíveis riscos diante de um procedimento cirúrgico.

Foram analisados conjunto de dados no período de 2009 a 2019, onde apresentaram 448 registros de pacientes que realizaram procedimentos, dos quais 247 tiveram ISC confirmada. Nos registros de amostras 269 são pacientes do gênero feminino e 179 masculino, com idades variando entre 1 e 97 anos, sendo em média pacientes com idade em torno de 45 anos. Outro fator interessante é que 216 apresentavam ao menos uma comorbidade, sendo 117 idosos, 311 adultos, 20 crianças e três cadeirantes. O Quadro 2 apresenta os grupos e as respectivas comorbidades.

Quadro 2 - Agrupamento das comorbidades

Grupo criado	**Comorbidades**
Cardiovasculares	• Arritmia • Cardiopatia • Hipertensão
Respiratórias	• Asma • DPOC • Bronquite
Autoimunes	• HIV • Doença de Crohn • Lúpus
Metabólicas	• Doença tireoideana
Psiquiátricas	• TOC • Transtorno psiquiátrico • Deficiência Cognitiva
Obesidade	• Obesidade • Obesidade mórbida
Diabetes	• Diabetes
Dependência química	• Alcoólatra • Usuário de drogas • Tabagista

Neoplasia	• Neoplasia
Outras	• Alzheimer • Anemia falciforme • Cirrose • Depressão • Dislipidemia • Diverticulite • Excesso de ferro • Hepatite • Hidrocefalia • Insuficiência renal • Insuficiência Renal Crônica • Paralisia de *Bell* • Psoríase • Trombofilia • Síndrome desconhecida

Fonte: Elaborado pelo AUTOR (2021).

De acordo com Quadro 3 é apresentado a Categorização etária:

Quadro 3 - Categorização etária

Classificação etária	**Regra**
Criança	Idade < 18
Adulto	Idade >= 18 & idade < 60
Idoso	Idade >= 60

Fonte: Elaborado pelo AUTOR (2021).

Exemplo de código genérico que faz o treinamento dos modelos representado na Figura 41:

Figura 41 – Código genérico para treinamento dos modelos.

```
def train_model(df, label, classifier):

    x = df.drop([label], axis = 1)
    y = df[label]

    X_train, X_test, y_train, y_test = split(x, y, test_size = TEST_SIZE, random_state = RANDOM_STATE)

    model = classifier.fit(X_train, y_train)

    # Train
    y_predicted = model.predict(X_train)
    plot_roc_curve(y_train, y_predicted, model)
    plot_metrics(y_train, y_predicted)

    # Testing
    y_predicted = model.predict(X_test)
    plot_roc_curve(y_test, y_predicted, model, 'Testing')
    plot_metrics(y_test, y_predicted)

    return model
```

Fonte: Elaborado pelo AUTOR (2021).

Este método recebe uma instância do algoritmo (*classifier*), realiza o split entre os dados de treinamento e teste. Faz o treinamento, calcula as métricas, imprime a curva ROC e a matriz de confusão. Por meio deste método, é garantido que todos os algoritmos irão passar pelos mesmos procedimentos e utilizarão os mesmos parâmetros básicos.

5.3 CONTRIBUIÇÕES DO TRABALHO

5.3.1 Robô Laura

Robô Laura foi criado por Jacson Fresatto, um analista de sistemas de Curitiba - PR e é considerado o primeiro desenvolvido para gerenciar riscos na área da saúde. Este projeto teve seus testes iniciais no Hospital Nossa Senhora das Graças, em Curitiba, com objetivo de reduzir o tempo de espera para inserir os dados dos pacientes de 3h42min para 42 minutos (KURTZ, 2017).

Conforme enfatiza Kurtz (2017), para o desenvolvimento do Robô Laura foi utilizada tecnologias de computação cognitiva e *machine learning* e entre suas finalidades está seu uso para analisar os bancos de dados de hospitais e outras instituições de saúde visando favorecer a eficiência dos atendimentos. No entanto, o principal objetivo do Robô Laura é a redução do tempo de identificação de caos de infecção generalizada, conhecida como sepse, e intensificar a velocidade de administrar antibióticos, que é fundamental para salvar a vida de um paciente. Neste sentido, no Brasil o tempo entre a suspeita e a confirmação do caso é de 13 horas, e com ajuda do Laura diminui para três horas.

Kurtz (2017), ainda acrescenta que, o sistema do Robô Laura funciona conversando diretamente com a área operacional e gerenciando riscos por meio da computação cognitiva, permitindo que o mesmo aprenda de acordo com as informações coletadas e se adaptando aos novos contextos. O Robô Laura utiliza terminais no hospital para realizar a comunicação com enfermeiros e médicos, e desta forma, quando um paciente necessita de atenção um alerta de urgência é mostrado. Em casos em que este pedido não é respondido, entra em ação uma ferramenta chamada Ansiedade de Laura, o qual deixa a cor do monitor cada vez mais vermelha, demonstrando desta forma, o aumento da urgência no caso. Em última alternativa, o robô entra em contato com os médicos responsáveis pelo paciente.

Com inicio da pandemia a aplicação do Robô Laura foi expandida para atender essa nova demanda e auxiliar e evitar a superlotação dos hospitais. Deste modo foi criado o Pronto Atendimento Digital (P.A Digital) que foi lançado em abril de 2020 e sua função é realizar uma triagem virtual. O Robô Laura realiza uma conversa por meio de um chatbot para analisar as respostas do usuário, identificando sintomas e alertando os casos suspeitos, orientando sem precisar sair de casa. O robô ainda acompanha o usuário/paciente por 14 dias por meio do *WhatsApp*, com monitoramente do quadro clínico, e caso o paciente tenha uma piora em seu estado, é recomendado que a mesma receba atendimento médico. Todos os procedimentos realizados pelo Robô Laura seguem parâmetros estabelecidos pela Organização Mundial da Saúde (OMS) e Ministério da Saúde (FANTINATO, 2021).

Para Mello (2019) as principais funções e especialidades do Robô Laura são descritos a seguir:

- Sepsis: é realizado o monitoramento pelo software em tempo real dos dados vitais dos pacientes para identificar antecipadamente a sepse. Os processos de tratamento e mediação da doença são acelerados com uso da tecnologia cognitiva;
- Compliance: para identificar falhas operacionais em uma organização, o robô realiza o monitoramento de processo e atividades, gerando alertas em casos em que as tarefas não estejam em conformidade com as boas práticas;
- *Epidemology:* auxilia o trabalho da vigilância epidemiológica com a monitoração de exames laboratoriais, identificando surtos e epidemias em tempo real;
- *Blood*: o robô realiza o monitoramento de bancos de sangue o que pode possibilitar a criação de campanhas de doação com objetivo de atingir doadores de um tipo específico de sangue. Além disso, prevê o perfil de risco para procedimentos cirúrgicos, detectando, em certos casos, a necessidade de transfusão.

- *Medicine*: realiza o monitoramento do comportamento das bactérias nos ambientes hospitalares com a finalidade de viabilizar um perfil de multirresistência. Também realiza o monitoramento de prescrições médicas e em casos de uso inadequado de antibióticos gera um alerta e controla a ingestão em pacientes críticos.
- *Saving*: os recursos das instituições são monitorados pelo robô, e caso exista um mau uso dos mesmos, pode oferecer uma visão estratégica de custos.

Além disso, como complementa Mello (2019), o Robô Laura pode possibilitar alguns benefícios aos hospitais, tais como:

- Otimização de processos: realiza a análise do fluxo de informações e dados cadastrados de pacientes para agilizar o trabalho dos profissionais de saúde. Desta forma, identifica falhas nas rotinas hospitalares para que a gestão treine sua equipe e melhore os processos;
- Redução de custos: realiza o monitoramente do uso de recursos do hospital para identificar erros, fazendo com que a administração tome atitudes para garantir a sustentabilidade financeira da instituição;
- Melhor atendimento ao paciente: o robô realiza o monitoramento do paciente e em casos em que ele identifique sinais de alerta, como mudança de temperatura ou parâmetros sanguíneos, entra em contato com as equipes para que as mesmas consigam agir prontamente. Em casos de sepse, o alerta pode evitar a sua evolução e, por consequência, evitar o óbito do paciente.

5.3.2 *IBM Watson Health*

Como acrescenta Mello (2019), a IBM (*International Business Machines*) é considerada uma das maiores empresas do mundo no âmbito da IA, e tem como um dos representantes o programa *Watson Health*. A plataforma tem foco na análise de dados na área da saúde e tem capacidade de processar e apresentar informações com exatidão, segurança e maciça para inúmeras empresas, profissionais e pacientes.

O *IBM Watson Health* é uma combinação dos principais pontos fortes da *IBM*, como experiência no setor da saúde, soluções avançadas de tecnologia com uso de IA, *blockchain*, dados e análises, entre outros, além de uma consultoria experiente, possibilita que o Watson ofereça suporte para as transformações digitais de vários clientes da empresa (IBM, 2021).

O lançamento do *Watson Health* ocorreu em 2015 e sua primeira versão era direcionada as instituições hospitalares. A partir de 2016, consultórios e clínicas do Brasil poderiam começar a serem clientes da *IBM* e utilizar o *Watson Health* (MELLO, 2019).

O principal objetivo do Watson Health, como descreve Mello (2019), desde o inicio foi fomentar a inovação com intuito de tratar das dificuldades dos problemas da saúde que necessitam de urgência a nível mundial. Neste sentido, com o uso da Inteligência Artificial o Watson Health coleta dados e insights na área da saúde para que possam ser utilizados em diferentes instituições. Além disso, a plataforma permite o desenvolvimento das próprias programações cognitivas, e desta forma, as empresas podem ser favorecidas com uso desta tecnologia. Um exemplo de aplicação do Watson Health é o uso de IA para coletar e analisar informações genéticas dos pacientes e cruzar essas informações com estudos científicos do mundo todo para estabelecer uma relação, por exemplo, com tipos diferentes de tumores para que o algoritmo apresente um resultado detalhado ao médico para que o mesmo consiga analisar o tratamento mais adequado com as terapias e medicamentos adequados conforme o diagnóstico.

Conforme completa Mello (2019), o Watson Health apresenta algumas subdivisões para diferentes especialidades:

Oncologia e genômica: *Watson Oncology* uma das áreas mais consultadas da ferramenta que tem por propósito aperfeiçoar o tratamento contra o câncer focado no paciente com metodologias e manuseio para dor apresentados são cada vez mais personalizados e mais precisos.

Ciências da vida: consiste em uma especialidade para acelerar terapias voltadas para atuarem de forma mais eficaz, ou seja, considera a premissa de que os estudos clínicos devem falhar mais rápido para que os recursos disponíveis para a área da saúde sejam gastos de modo mais rápido e coerente.

Valor baseado no cuidado: está focado em reduzir gastos e entregar qualidade. Leva em consideração os passos que devem ser seguidos para promover tanto o sistema em questão como também as questões operacionais e financeiras do cenário mundial.

Ações governamentais: tem o objetivo de potencializar programas governamentais, principalmente, para países em desenvolvimento, como o Brasil. Ou seja, melhorar o valor agregado na saúde abrange a promoção de serviços sociais e iniciativas para facilitar a qualidade de vida dos cidadãos mais desamparados.

Entre os principais benefícios proporcionados pelo *Watson Health* está o auxílio em tomadas de decisão até provedores de saúde e equipes de cuidado assistencial para que o trabalho seja realizado de forma mais integrada e efetiva. Isto porque a ferramenta oferece aos médicos o acesso ao histórico completo de seus pacientes, complicações indesejadas e desta forma, gastos com saúde podem ser eficientemente reduzidos. Bem como, diagnósticos mais precisos podem gerar economia para o sistema de saúde.

5.3.3 Discussão dos resultados em comparação ao Robô Laura e *Watson Health*

Ao realizar uma comparação resgatando e comparando projetos próximos ao Sistema que está sendo apresentado, pode-se analisar o Robô Laura, criado por Jacson Fresatto, um sistema desenvolvido com tecnologias de computação cognitiva e *machine learning*, que utiliza banco de dados que visa favorecer a eficiência de atendimentos e gerenciar riscos. Porém, seu principal objetivo é identificar em menor tempo possível focos de infecção generalizada ou sepse, intensificando os procedimentos cabíveis como medicação no intuito de salvar a vida do paciente. Do contrário, o sistema proposto analisa e cruza informações enviando sinal de alerta antes mesmo de se consolidar a infecção, antecipando possibilidades fazendo com que a equipe médica se posicione preventivamente diante do alerta.

O Robô Laura, assim como no sistema proposto, utiliza terminais no hospital para realizar a comunicação com enfermeiros e médicos. E tem por resultado final uma redução de custos, da mesma forma a que este vem apresentar.

Diferente do estudo atual, *Watson*, lançado em 2015 utiliza de uma plataforma limitada em análise de dados apresentando informações exatas e seguras, tanto para profissionais quanto para pacientes, igualmente utiliza da tecnologia de IA, *blockchain*, dados e análise. Tem por objetivo tratar dos problemas de saúde de urgência, que através da IA, coleta de dados e *insights*, analisa informações genéticas e cruzar com estudos científicos para estabelecer uma relação com tipos diferentes de tumores para que o algoritmo apresente um resultado detalhado ao médico para que o mesmo consiga analisar o tratamento mais adequado com as terapias e medicamentos adequados conforme o diagnóstico.

Uma das áreas mais consultadas é a Oncologia e Genômica podendo assim aperfeiçoar o tratamento do paciente com o uso de metodologias cada vez mais precisas. *Watson* se destaca pelo auxílio na tomada de decisão dos profissionais, pois a ferramenta

oferece aos médicos o acesso ao histórico completo de seus pacientes, complicações indesejadas e desta forma, gastos com saúde podem ser eficientemente reduzidos.

6 CONCLUSÃO

Partindo do levantamento de dados e informações, desenvolvimento e uma série de testes, no sistema proposto percebe-se um impacto positivo no âmbito hospitalar onde a pesquisa foi realizada. Realizou-se ainda pesquisa bibliográfica para contextualizar o trabalho, que através da literatura e a escolha de técnicas e programas fizeram com que o sistema atingisse a proposta central de aplicação.

É sabido que as Tecnologias de Informação e Comunicação estão em constante evolução, presentes cada vez com maior frequência, e diante do cenário hospitalar atual fica claro perceber o índice considerável de risco de infecção pós-cirúrgica, onde traz consigo uma série de preocupações e problemas tanto para o hospital quanto para o paciente.

A finalidade da construção dessa tecnologia foi contribuir para a redução da incidência das infecções nosocomiais, por meio da disponibilização de um instrumento que avalie precocemente os riscos que o paciente apresenta segundo as comorbidades prévias existentes no doente.

Sendo assim, a Inteligência Artificial juntamente com a Mineração de dados são tecnologias que vem propor meios de identificar preventivamente possíveis infecções diante de um procedimento cirúrgico, haja vista que as graves consequências impostas aos pacientes que desenvolveram a ISC determinam a necessidade de dedicar esforços para a criação de estratégias para a prevenção dessa infecção. Neste sentido, a identificação dos fatores de risco para a ISC contribui para a adoção precoce de intervenções de enfermagem que objetivam minimizar esse tipo de complicação pós-operatória.

Sendo o intuito principal deste trabalho o desenvolvimento de um sistema que irá coletar e analisar informações, com o apoio de técnicas de mineração de dados e inteligência artificial percebemos que em um ambiente hospitalar, a possibilidade de automatizar este processo, representaria um avanço tecnológico significativo, trazendo como resultados a prevenção desta classe de síndrome infecciosa e consigo reflexos positivos para o paciente e para o hospital evitando uma série de problemas e transtornos como, aumento de estadia hospitalar, que aumentam as chances de novas infecções, risco de novos procedimentos cirúrgicos e risco de óbito, do outro lado um alto custo para tratar e recuperar a vida do paciente com a utilização de medicações, leito, profissionais qualificados, gerando assim, um custo que se pudesse ser evitado, teria como investir em melhorias com tecnologia e equipamentos hospitalares.

Por fim, sugere-se para trabalhos futuros que seja avaliado em setores diferentes da instituição hospitalar para que não se resuma em apenas avaliação de Infecções do Sítio Cirúrgico, mas em todas as possibilidades de infecções no âmbito Hospitalar, podendo ainda analisar modelos mais avançados de Deep Learning, incluir um histórico completo com mais dados dos pacientes e exames, afinal quanto mais informações mais assertivo fica o resultado final do sistema, e porque não ter um acesso entre unidades de saúde, podendo assim cruzar dados e informações enriquecendo ainda mais o prontuário de cada paciente. Outra etapa um tanto futurista e desafiadora é sugerir tentar através de projetos de sistema como esse prever a severidade da infecção e porque não o óbito do paciente, a tecnologia é ilimitada e desenvolver programações que disponibilizam de resultados que possam definir uma alternativa de mudança através da antecipação para então se obter de uma nova postura no sistema da saúde.

REFERÊNCIAS

ADLASSNIG, Klaus-Peter; BLACKY, Alexander; KOLLER, Walter. Artificial-Intelligence-Based Hospital-Acquired Infection Control. **Studies In Health Technology And Informatics**, [S.L.], v. 149, n. , p. 103-110, 2009. IOS Press. http://dx.doi.org/10.3233/978-1-60750-050-6-103. Disponível em: http://www.meduniwien.ac.at/kpa/publications/2009_Bushko_SHTI149_--_AI-Based_Hospital-Acquired_Infection_Control.pdf. Acesso em: 21 abr. 2020.

AMR, Tarek. **Hands-On Machine Learning with scikit-learn and Scientific Python Toolkits**: a practical guide to implementing supervised and unsupervised machine learning algorithms in python. Birmingham, UK: Packt Publishing Ltd., 2020. 368 p.

BAGCI, Ulas *et al*. Automatic Detection and Quantification of Tree-in-Bud (TIB) Opacities From CT Scans. **IEEE Transactions On Biomedical Engineering**, [S.L.], v. 59, n. 6, p. 1620-1632, jun. 2012. Institute of Electrical and Electronics Engineers (IEEE). http://dx.doi.org/10.1109/tbme.2012.2190984. Disponível em: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3511590/. Acesso em: 17 ago. 2020.

BANAEE, Hadi; AHMED, Mobyen; LOUTFI, Amy. Data Mining for Wearable Sensors in Health Monitoring Systems: a review of recent trends and challenges. **Sensors**, [S.L.], v. 13, n. 12, p. 17472-17500, 17 dez. 2013. MDPI AG. http://dx.doi.org/10.3390/s131217472. Disponível em: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3892855/. Acesso em: 10 maio 2021.

BAYSARI, Melissa T. *et al*. The effectiveness of information technology to improve antimicrobial prescribing in hospitals: a systematic review and meta-analysis. **International Journal Of Medical Informatics**, [S.L.], v. 92, p. 15-34, ago. 2016. Elsevier BV. http://dx.doi.org/10.1016/j.ijmedinf.2016.04.008. Disponível em: https://europepmc.org/article/med/27318068. Acesso em: 15 ago. 2020.

BERNARD, Zoë. **So, what is machine learning anyways?**: here's a quick breakdown. Here's a quick breakdown. 2017. Business Insider. Disponível em: https://www.businessinsider.com/what-is-machine-learning-quick-explainer-2017-11. Acesso em: 11 out. 2020.

BOXWALA, Aziz A. *et al*. Toward a Representation Format for Sharable Clinical Guidelines. **Journal of Biomedical Informatics**, [S.L.], v. 34, n. 3, p. 157-169, jun. 2001. Elsevier BV. http://dx.doi.org/10.1006/jbin.2001.1019.

BRAGA, Ana Vitória *et al*. INTELIGÊNCIA ARTIFICIAL NA MEDICINA. **III CIPEEX**: Ciência para a redução das desigualdades, Anápolis, p. 937-941, 2018. XV Mostra de Saúde - 10 anos do Curso de Medicina. Disponível em: http://anais.unievangelica.edu.br/index.php/CIPEEX/article/view/2997/1348. Acesso em: 22 abr. 2021.

BRANTS, Thorsten *et al*. Large Language Models in Machine Translation. **Proceedings Of The 2007 Joint Conference On Empirical Methods In Natural Language Processing And Computational Natural Language Learning (Emnlp-Conll)**, Prague, Czech Republic, p. 858-867, 2007. Disponível em: https://www.aclweb.org/anthology/D07-1090/. Acesso em: 24 nov. 2020.

BRASIL. Constituição (1997). Lei nº 9.431, de 6 de janeiro de 1997. **Dispõe sobre a obrigatoriedade da manutenção de Programa de Controle de Infecções Hospitalares pelos Hospitais do País**. Brasília, DF: Presidência da República. Casa Civil. Subchefia Para Assuntos Jurídicos., 6 jan. 1997. Disponível em: http://www.planalto.gov.br/ccivil_03/leis/l9431.htm#:~:text=LEI%20N%C2%BA%209.431%2C%20DE%206%20DE%20JANEIRO%20DE%201997.&text=Disp%C3%B5e%20sobre%20a%20obrigatoriedade%20da,Art. Acesso em: 12 out. 2020.

BRASIL. Nacional de Vigilância Sanitária, ANVISA. **Investigação e controle de bactérias multirresistentes**. Brasília – DF: Agência Nacional de Vigilância Sanitária, 2007. 21 p. Gerência de Investigação e Prevenção das Infecções e dos Eventos Adversos (GIPEA)/Gerência Geral de Tecnologia em Serviços de Saúde (GGTES). Disponível em: https://www.anvisa.gov.br/servicosaude/controle/reniss/manual%20_controle_bacterias.pdf. Acesso em: 26 ago. 2020.

BRASIL. Agência Nacional de Vigilância Sanitária, ANVISA. **Critérios Diagnósticos de Infecções Relacionadas à Assistência à Saúde**. Brasília – DF: Agência Nacional de Vigilância Sanitária, 2013. 84 p. (Segurança do Paciente e Qualidade em Serviços de Saúde). Disponível em: http://bvsms.saude.gov.br/bvs/publicacoes/criterios_diagnosticos_infeccoes_assistencia_saude.pdf. Acesso em: 05 out. 2020.

BRASIL. Nacional de Vigilância Sanitária, ANVISA. **Medidas de Prevenção de Infecção Relacionada à Assistência à Saúde**. Brasília – DF: Agência Nacional de Vigilância Sanitária, 2017. 122 p. (Segurança do Paciente e Qualidade em Serviços de Saúde). Gerência de Vigilância e Monitoramento em Serviços de Saúde (GVIMS). Gerência Geral de Tecnologia em Serviços de Saúde (GGTES). Disponível em: http://www.riocomsaude.rj.gov.br/Publico/MostrarArquivo.aspx?C=pCiWUy84%2BR0%3D. Acesso em: 15 out. 2020.

BROSSETTE, Stephen E.; HYMEL, Patrick A.. Data Mining and Infection Control. **Clinics In Laboratory Medicine**, [S.L.], v. 28, n. 1, p. 119-126, mar. 2008. Elsevier BV. http://dx.doi.org/10.1016/j.cll.2007.10.007. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0272271207001163. Acesso em: 30 maio 2021.

CALEGARI, Ana Paula Katz. **Medicina 4.0**: um brinde à saúde na era da inteligência artificial. 2019. DWIH São Paulo. Disponível em: https://www.dwih-saopaulo.org/pt/2019/05/06/medicina-4-0-um-brinde-a-saude-na-era-da-inteligencia-artificial/. Acesso em: 17 maio 2021.

CÁNOVAS-SEGURA, Bernardo; CAMPOS, Manuel; MORALES, Antonio; JUAREZ, Jose M.; PALACIOS, Francisco. Development of a clinical decision support system for antibiotic management in a hospital environment. **Progress In Artificial Intelligence**, [S.L.], v. 5, n. 3, p. 181-197, 3 mar. 2016. Springer Science and Business Media LLC. http://dx.doi.org/10.1007/s13748-016-0089-x. Disponível em: https://www.semanticscholar.org/paper/Development-of-a-clinical-decision-support-system-a-C%C3%A1novas-Segura-Campos/adf3782afb32a74059b96a42547d70cd31685f10. Acesso em: 23 maio 2021.

CARVALHO, Deborah Ribeiro; ESCOBAR, Leandro Fabian Almeida; TSUNODA, Denise. Pontos de Atenção para o Uso da Mineração de Dados da Saúde. **Informação & Informação**, [S.L.], v. 19, n. 1, p. 249-273, 23 fev. 2014. Universidade Estadual de Londrina. http://dx.doi.org/10.5433/1981-8920.2014v19n1p249. Disponível em: https://www.researchgate.net/publication/262971356_Pontos_de_Atencao_para_o_Uso_da_Mineracao_de_Dados_da_Saude. Acesso em: 10 maio 2021.

CARVALHO, Rafael Lima Rodrigues de *et al*. Incidence and risk factors for surgical site infection in general surgeries. **Revista Latino-Americana de Enfermagem**, [S.L.], v. 25, p. 1-8, 4 dez. 2017. FapUNIFESP (SciELO). http://dx.doi.org/10.1590/1518-8345.1502.2848. Disponível em: https://www.scielo.br/scielo.php?pid=S0104-11692017000100390&script=sci_abstract&tlng=pt. Acesso em: 10 ago. 2020.

CARVALHO, Ricardo César de. Aplicação de mineração de dados em informações oriundas de prontuários de paciente. **Informação em Pauta**, [S.L.], p. 161-181, 26 nov. 2018. Portal de Periódicos da UFC. http://dx.doi.org/10.32810/2525-3468.ip.v3iespecial.2018.39723.161-181. Disponível em: http://repositorio.ufc.br/bitstream/riufc/38098/1/2018_art_rccarvalho.pdf. Acesso em: 22 abr. 2021.

CASALI, Fabiana Tambellini *et al*. Análisis de las características epidemiológicas de la fiebre amarilla en un estado del sureste de Brasil. **REVENF**: Revista Enfermería Actual en Costa Rica, Costa Rica, p. 1-16, dez. 2019. Semestral. Universidad de Costa Rica. Disponível em: https://www.scielo.sa.cr/pdf/enfermeria/n37/1409-4568-enfermeria-37-50.pdf. Acesso em: 14 nov. 2020.

CDC, Centers For Disease Control And Prevention. **Healthcare-associated Infections (HAI)**: types of healthcare-associated infections. Types of Healthcare-associated Infections. 2014. Centers for Disease Control and Prevention, National Center for Emerging and Zoonotic Infectious Diseases (NCEZID), Division of Healthcare Quality Promotion (DHQP). Disponível em: https://www.cdc.gov/hai/infectiontypes.html. Acesso em: 31 jan. 2020.

CDC, Centers For Disease Control And Prevention. **Healthcare-associated Infections (HAI)**: types of infections. Surgical Site Infection (SSI). 2010. Centers for Disease Control and Prevention, National Center for Emerging and Zoonotic Infectious Diseases (NCEZID), Division of Healthcare Quality Promotion (DHQP). Disponível em: https://www.cdc.gov/hai/ssi/ssi.html. Acesso em: 02 nov. 2020.

COHEN, Jacob *et al*. Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences. **Lawrence Erlbaum Associates, Publishers**, [S.L.], p. 1-73, 17 jun. 2013. Routledge. http://dx.doi.org/10.4324/9780203774441.

COELHO, Lucas. **Machine Learning**: o que é, conceito e definição. O que é, conceito e definição. 2020. Data Analytics, Big Data, Data Science – Blog Cetax. Artigos, materiais e tutoriais de Business Intelligence, Big Data, Data Warehouse e ETL. Disponível em: https://www.cetax.com.br/blog/machine-learning/. Acesso em: 17 ago. 2020.

CRESWELL, John W.. **Projeto de pesquisa**: métodos qualitativo, quantitativo e misto. 2. ed. Porto Alegre: Artmed, 2007. 248 p.

CROSS, S.E; WALKER, E.. DART: applying knowledge based planning and scheduling to crisis action planning. In: ZWEBEN, M.; FOX, M. S.. **Intelligent Scheduling**. San Francisco: Morgan Kaufmann, 1994. p. 711-729.

DASHGOO. **Machine Learning**: o que é e por que é importante. O Que É e Por Que é Importante. 2018. Marketing Digital. Disponível em: https://dashgoo.com/machine-learning-o-que-e-e-por-que-e-importante/. Acesso em: 12 out. 2020.

DOMINGOS, Pedro. A few useful things to know about machine learning. **Communications of the ACM**, [S.L.], v. 55, n. 10, p. 78-87, out. 2012. Association for Computing Machinery (ACM). http://dx.doi.org/10.1145/2347736.2347755.

FANTINATO, Giovanna. **Inteligência Artificial brasileira ajuda a combater covid-19**. 2021. Tecmundo. Disponível em: https://www.tecmundo.com.br/ciencia/215038-robo-laura-ia-ajudando-diagnosticos-covid-19.htm. Acesso em: 27 jun. 2021.

FAYYAD, Usama; PIATETSKY-SHAPIRO, Gregory; SMYTH, Padhraic. From Data Mining to Knowledge Discovery in Databases. **Ai Magazine**, [S. L.], v. 17, n. 3, p. 37-54, 1996. Disponível em: https://ojs.aaai.org//index.php/aimagazine/article/view/1230/. Acesso em: 22 maio 2021.

FERRAZ, Álvaro Antonio Bandeira *et al*. Infecção de sítio cirúrgico após cirurgia bariátrica: resultados de uma abordagem com pacote de cuidados.. **Revista do Colégio Brasileiro de Cirurgiões**, [S.L.], v. 46, n. 4, p. 1-8, 2019. FapUNIFESP (SciELO). http://dx.doi.org/10.1590/0100-6991e-20192252. Disponível em: https://www.scielo.br/scielo.php?pid=S0100-69912019000400153&script=sci_arttext&tlng=en. Acesso em: 17 ago. 2020.

FITZPATRICK, Fidelma; DOHERTY, Aaron; LACEY, Gerard. Using Artificial Intelligence in Infection Prevention. **Current Treatment Options In Infectious Diseases**, [S.L.], v. 12, n. 2, p. 135-144, 19 mar. 2020. Springer Science and Business Media LLC. http://dx.doi.org/10.1007/s40506-020-00216-7. Disponível em: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7095094/. Acesso em: 15 jun. 2020.

FORATO, Fidel. **HACKMED**: como a ia está transformando a medicina brasileira. Como a IA está transformando a medicina brasileira. 2020. Canaltech. Disponível em: https://canaltech.com.br/saude/hackmed-como-a-ia-esta-mudando-a-medicina-brasileira-159969/. Acesso em: 01 jun. 2021.

FORTUNA, Fernanda. **Como o Watson está ajudando o Fleury na medicina de precisão**. 2018. Saúde Business. Disponível em: https://www.saudebusiness.com/ti-e-inovao/como-o-watson-est-ajudando-o-fleury-na-medicina-de-preciso. Acesso em: 29 jun. 2021.

GANDHI, Priyanka; TANDON, Neelam. Application of Web Data Mining Techniques in CRM for Its Support to Health Industry. **Ssrn Electronic Journal**, [S.L.], p. 1-6, 2019. Elsevier BV. http://dx.doi.org/10.2139/ssrn.3446619. Disponível em: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3446619. Acesso em: 23 abr. 2021

GARLA, Vijay N.; BRANDT, Cynthia. Ontology-guided feature engineering for clinical text classification. **Journal Of Biomedical Informatics**, [S.L.], v. 45, n. 5, p. 992-998, out. 2012. Elsevier BV. http://dx.doi.org/10.1016/j.jbi.2012.04.010. Disponível em: https://www.sciencedirect.com/science/article/pii/S1532046412000639. Acesso em: 6 nov. 2020.

GÉRON, Aurélien. **Hands-on Machine Learning with Scikit-Learn, Keras & TensorFlow**: concepts, tools, and techniques to build intelligent systems. 2. ed. Sebastopol, CA: O'Reilly Media, Inc, 2019. 851 p.

GEURTS, Pierre; ERNST, Damien; WEHENKEL, Louis. Extremely randomized trees. **Machine Learning**, [S.L.], v. 63, n. 1, p. 3-42, 2 mar. 2006. Springer Science and Business Media LLC. http://dx.doi.org/10.1007/s10994-006-6226-1.

GIL, Antonio Carlos. **Como elaborar projetos de pesquisa.** 5. ed. São Paulo: Atlas, 2010. 200 p.

GOODMAN, David; KEENE, Raymond. **Man versus Machine**: Kasparov versus Deep Blue. Manassas: H3 Inc., 1997. 128 p.

GOODMAN, Joshua; HECKERMAN, David. Fighting spam with statistics. **Significance**, [S.L.], v. 1, n. 2, p. 69-72, 26 maio 2004. Wiley. http://dx.doi.org/10.1111/j.1740-9713.2004.021.x. Disponível em: https://rss.onlinelibrary.wiley.com/doi/full/10.1111/j.1740-9713.2004.021.x. Acesso em: 10 nov. 2020.

GÓMEZ-GONZÁLEZ, Emilio *et al*. Artificial intelligence in medicine and healthcare: a review and classification of current and near-future applications and their ethical and social impact. **ArXiv**, New York, p. 1-20, 2020. Cornell University - Computers and Society. Disponível em: https://arxiv.org/abs/2001.09778v2. Acesso em: 15 jun. 2020.

GRANVILLE, Lisandro Zambenedetti. **Machine Learning**: desafios para um Brasil competitivo. Porto Alegre: Revista Da Sociedade Brasileira de Computação, 2019. 46 p. Disponível em: https://www.sbc.org.br/images/flippingbook/computacaobrasil/computa_39/pdf/CompBrasil_39_180.pdf. Acesso em: 12 jun. 2021.

GUPTA, Itisha; NAGPAL, Garima. **Artificial Intelligence and Expert Systems**. Dulles, VA: Mercury Learning And Information LLC, 2020. 424 p.

HACKELING, Gavin. **Mastering Machine Learning with scikit-learn**: learning algorithms to real-world problems using scikit-learn. 2. ed. Birmingham, UK: Packt Publishing Ltd., 2017. 249 p.

HAN, Jiawei; KAMBER, Micheline; PEI, Jian. **Data Mining**: concepts and techniques. 3. ed. Burlington, Massachusetts: Elsevier, 2012. 740 p.

HAND, David J.; MANNILA, Heikki; SMYTH, Padhraic. **Principles of Data Mining**. London: MIT Press, 2001. 578 p.

HORAN, Teresa C.; ANDRUS, Mary; DUDECK, Margaret A.. CDC/NHSN surveillance definition of health care–associated infection and criteria for specific types of infections in the acute care setting. **American Journal Of Infection Control**, [S.L.], v. 36, n. 5, p. 309-332, jun. 2008. Elsevier BV. http://dx.doi.org/10.1016/j.ajic.2008.03.002. Disponível em: https://www.ajicjournal.org/article/S0196-6553(08)00167-3/fulltext. Acesso em: 10 nov. 2020.

IBM. **O que é o IBM Watson Health?**: o IBM Watson Health está empenhado em ajudar a criar ecossistemas de saúde mais inteligentes.. O IBM Watson Health está empenhado em ajudar a criar ecossistemas de saúde mais inteligentes.. 2021. O Watson Health é a saúde mais inteligente. Disponível em: https://www.ibm.com/br-pt/watson-health. Acesso em: 29 jun. 2021.

IZENMAN, Alan Julian. **Modern Multivariate Statistical Techniques**: regression, classification, and manifold learning. Philadelphia, PA: Springer, 2008. 757 p. Springer Texts in Statistics.

JACCARD, James; TURRISI, Robert; WAN, Choi K. Interaction effects in multiple regression. **Sage University Papers Series.**: Quantitative applications in the social sciences, Newbury Park, v. 7, n. 72, p. 1-95, 1990.

JANERT, Philipp K.. **Data Analysis with Open Source Tools**. *Sebastopol, CA:* O'Reilly Media, Inc, 2011. 533 p.

KAUARK, Fabiana da Silva; MANHÃES, Fernanda Castro; SOUZA, Carlos Henrique Medeiros de. **Metodologia Da Pesquisa**: um guia pratico. Itabuna: Via Litterarum, 2010. 96 p.

KURTZ, João. **Conheça o Robô Laura, inovação brasileira na saúde para salvar vidas**. 2017. TechTudo. Disponível em: https://www.techtudo.com.br/noticias/2017/12/conheca-o-robo-laura-inovacao-brasileira-na-saude-para-salvar-vidas.ghtml. Acesso em: 26 jun. 2021.

LAMMA, E. *et al*. Artificial Intelligence Techniques for Monitoring Dangerous Infections. **IEEE Transactions On Information Technology In Biomedicine**, [S.L.], v. 10, n. 1, p. 143-155, jan. 2006. Institute of Electrical and Electronics Engineers (IEEE). http://dx.doi.org/10.1109/titb.2005.855537. Disponível em: https://www.researchgate.net/publication/3415804_Artificial_Intelligence_Techniques_for_Monitoring_Dangerous_Infections. Acesso em: 10 abr. 2020.

LARSON, Elaine. Innovations in Health Care: antisepsis as a case study. **Public Health**: Then and Now, S.I, v. 79, n. 1, p. 92-99, 1989. Disponível em: https://ajph.aphapublications.org/doi/pdf/10.2105/AJPH.79.1.92. Acesso em: 10 nov. 2020.

LINOFF, Gordon S.; BERRY, Michael J. A. **Data Mining Techniques**: For marketing sales and customer relationship management. 3.ed. John Wiley & Sons: 2011. 847p.

LOBO, Luiz Carlos. Inteligência Artificial e Medicina. **Revista Brasileira de Educação Médica**, [S.L.], v. 41, n. 2, p. 185-193, jun. 2017. FapUNIFESP (SciELO). http://dx.doi.org/10.1590/1981-52712015v41n2esp. Disponível em: https://www.scielo.br/scielo.php?script=sci_abstract&pid=S0100-55022017000200185&lng=pt&nrm=iso. Acesso em: 10 abr. 2020.

LOBO, Luiz Carlos. Inteligência artificial, o Futuro da Medicina e a Educação Médica. **Revista Brasileira de Educação Médica**, [S.L.], v. 42, n. 3, p. 3-8, set. 2018. FapUNIFESP (SciELO). http://dx.doi.org/10.1590/1981-52712015v42n3rb20180115editorial1. Disponível em: https://www.scielo.br/pdf/rbem/v42n3/1981-5271-rbem-42-3-0003.pdf. Acesso em: 22 abr. 2021.

LOPES, Gesiel Rios *et al*. Introdução à Análise Exploratória de Dados com Python. In: CONFERENCE: ESCOLA REGIONAL DE COMPUTAÇÃO APLICADA À SAÚDE (ERCAS), 8., 2019, Teresina, Pi. **Anais [...] .** Teresina: (Ercas, 2019. p. 1-18. Disponível em: https://www.researchgate.net/publication/336778766_Introducao_a_Analise_Exploratoria_de_Dados_com_Python. Acesso em: 23 abr. 2020.

MAGALHÃES, Péricles; SPÍNOLA, Rodrigo O.. SQL Magazine **– Mineração de Dados:** Tarefas e Técnicas, Rio de Janeiro: Devmedia, 123ª ed., 2014.

MALIK, Shubham; HARODE, Rohan; KUNWAR, Akash Singh. XGBoost: a deep dive into boosting (introduction documentation). **Technical Report**, S.I, p. 3-22, 2020. Disponível em: https://www.researchgate.net/publication/339499154. Acesso em: 17 abr. 2020.

MARASCUILO, Leonard A.; SERLIN, Ronald C.. **Statistical Methods for the Social and Behavioral Sciences**. New York: Freenan, 1988. 885 p.

MARCONI, M. de A.; LAKATOS, Eva Maria. **Fundamentos de metodologia científica**. São Paulo: Atlas, 2007. 368p.

MARGARIDO, Inês Tomás Marques Martins. **A medicina do futuro nos dias de hoje**: aplicações da inteligência artificial. 2020. 34 f. Dissertação (Mestrado) - Curso de Medicina, Fisiologia, Faculdade de Medicina de Lisboa, Lisboa, 2020. Disponível em: https://repositorio.ul.pt/handle/10451/46808. Acesso em: 14 maio 2021.

MCKINNEY, Wes. **Python for Data Analysis**: Data Wrangling with Pandas, Numpy, and IPython. 2. ed. Sebastopol, Ca: O'reilly Media, Inc., 2018. 541 p.

MELLO, Heloisa C.. **Robô Laura**: conheça mais essa inovação na área da saúde. conheça mais essa inovação na área da saúde. 2019. Medicalway. Disponível em: https://blog.medicalway.com.br/marco-19-robo-laura-conheca-mais-essa-inovacao-na-area-da-saude/. Acesso em: 27 jun. 2021.

MELLO, Heloisa C.. **Watson Health**: conheça esse programa e seus benefícios. conheça esse programa e seus benefícios. 2021. Medicalway. Disponível em: https://blog.medicalway.com.br/watson-health-conheca-esse-programa-e-seus-beneficios/. Acesso em: 29 jun. 2021.

MENDENHALL, William; BEAVER, Robert J.; BEAVER, Barbara M.. **Introduction to Probability and Statistics**. 15. ed. Sebastopol, CA: Cengage Learning, 2019. 744 p.

MENDES, Francisco Coêlho. **Administração de Sistemas de Informação**. Rio de Janeiro: Fundação CECIERJ, 2009. 190 p. Consórcio CEDERJ. Disponível em: https://canal.cecierj.edu.br/012016/48f9811be97600b2e56e0d1154140c18.pdf. Acesso em: 17 abr. 2020.

MITCHELL, Tom M.; MITCHELL, Thomas; THOMAS, Mitchell. **Machine Learning**. New York: McGraw-Hill Science/Engineering/Mat, 1997. 432 p.

ML, Rubix. **Extra Tree Classifier#**. 2021. Material for MkDocs. Disponível em: https://docs.rubixml.com/latest/classifiers/extra-tree-classifier.html. Acesso em: 18 jan. 2021.

MUJTABA, Hussain. **Hyperparameter Tuning with GridSearchCV**. 2020. Great Learning. Disponível em: https://www.mygreatlearning.com/blog/gridsearchcv/. Acesso em: 10 dez. 2020.

MUKHERJEE, Siddhartha. **AI versus MD**: what happens when diagnosis is automated. What happens when diagnosis is automated. 2017. The New Yorker. Disponível em: https://www.newyorker.com/magazine/2017/04/03/ai-versus-md. Acesso em: 15 abr. 2020.

NILSSON, Nils J.. **Introductiontomachine learning**: an early draft of a proposed textbook. Stanford, Ca: Department Of Computer Science. Stanford University, 1998. 188 p.

PADOVEZE, Maria Clara; FORTALEZA, Carlos Magno Castelo Branco. Healthcare-associated infections: challenges to public health in brazil. **Revista de Saúde Pública**, [S.L.], v. 48, n. 6, p. 995-1001, dez. 2014. FapUNIFESP (SciELO). http://dx.doi.org/10.1590/s0034-8910.2014048004825. Disponível em: https://www.scielo.br/pdf/rsp/v48n6/pt_0034-8910-rsp-48-6-0995.pdf. Acesso em: 15 nov. 2020.

PANSONATO, Ramon; TOMAZELA, Maria das Graças J. M. **Análise comparativa dos algoritmos de clusterização K-means e fuzzy c-fuzzy com uso de dados oriundos do plantio de cana-de-açucar**. Revista de Estudos e Reflexões Tecnológicas, Indaiatuba, n. 12, 2014.

PECK, Roxy; OLSEN, Chris; DEVORE, Jay L.. **Introduction to Statistics and Data Analysis**. 5. ed. Boston: Cengage Learning, 2016. 842 p.

PEDHAZUR, Elazar J.. **Multiple regression in behavioral research**: xplanation and prediction / elazar j. pedhazur. 2. ed. Fort Worth: Harcourt Brace College Publishers, 1982. 822 p.

PITTET, D. *et al*. Infection control as a major World Health Organization priority for developing countries. **Journal Of Hospital Infection**, [S.L.], v. 68, n. 4, p. 285-292, abr. 2008. Elsevier BV. http://dx.doi.org/10.1016/j.jhin.2007.12.013. Disponível em: https://www.journalofhospitalinfection.com/article/S0195-6701(08)00011-X/fulltext. Acesso em: 20 nov. 2020.

PRATES, Cassiana Gil *et al*. Comparação das taxas de infecção cirúrgica após implantação do checklist de segurança. **Acta Paulista de Enfermagem**, [S.L.], v. 31, n. 2, p. 116-122, mar. 2018. FapUNIFESP (SciELO). http://dx.doi.org/10.1590/1982-0194201800018. Disponível em: https://www.scielo.br/scielo.php?script=sci_arttext&pid=S0103-21002018000200116&lng=pt&tlng=pt. Acesso em: 17 ago. 2020.

PRODANOV, Cleber Cristiano; FREITAS, Ernani Cesar De. **Metodologia do trabalho científico**: Métodos e Técnicas da Pesquisa e do Trabalho Acadêmico. 2 ed. Novo Hamburgo: Feevale, 2013. 277 p.

RABELO, Agnes. **Machine Learning**: o que é e qual sua influência no marketing digital?. o que é e qual sua influência no marketing digital?. 2018. Rock Content. Disponível em: https://rockcontent.com/br/blog/machine-learning/. Acesso em: 12 out. 2020.

RAMESH, An *et al*. Artificial intelligence in medicine. **Annals Of The Royal College Of Surgeons Of England**, [S.L.], v. 86, n. 5, p. 334-338, 1 set. 2004. Royal College of Surgeons of England. http://dx.doi.org/10.1308/147870804290. Disponível em: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1964229/. Acesso em: 23 abr. 2021.

RASCHKA, Sebastian; MIRJALILI, Vahid. **Python Machine Learning**: machine learning and deep learning with python, scikit-learn, and tensorflow. 2. ed. Birmingham, UK: Packt Publishing Ltd., 2017. 623 p.

RIBEIRO, Wandy. **USP e Einstein usam inteligência artificial para diagnosticar coronavírus**. 2020. Postado em Farmácia Hospitalar. Disponível em: https://www.ictq.com.br/farmacia-hospitalar/1389-usp-e-einstein-usam-inteligencia-artificial-para-diagnosticar-coronavirus. Acesso em: 17 maio 2021.

ROUHIAINEN, Lasse. **Artificial Intelligence**: 101 things you must know today about our future. [S. L.]: Createspace Independent Publishing Platform, 2018. 300 p.

RUBIN, Michael A. *et al.* An Agent-Based Model for Evaluating Surveillance Methods for Catheter-Related Bloodstream Infection. **Amia Symposium Proceedings**, [S. L.], p. 631-635, 2008. Disponível em: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2655959/. Acesso em: 15 maio 2020.

RUSSELL, Stuart J.; NORVIG, Peter. **Inteligência Artificial**. 2. ed. Rio de Janeiro: Elsevier, 2004. 1040 p.

RUSSELL, Stuart J.; NORVIG, Peter. **Artificial Intelligence**: a modern approach. 3. ed. New Jersey: Pearson, 2010. 1152 p.

RUSSELL, Stuart J.; NORVIG, Peter. **Inteligência Artificial**. 3. ed. Rio de Janeiro: Elsevier, 2013. 1320 p.

SAGE, Andrew P.. **Concise Encyclopedia of Information Processing in Systems & Organizations**: advances in systems, control and information engineering. Oxford: Pergamon Pr, 1990. 548 p.

SAMWALD, Matthias *et al*. The Arden Syntax standard for clinical decision support: experiences and directions. **Journal Of Biomedical Informatics**, [S.L.], v. 45, n. 4, p. 711-718, ago. 2012. Elsevier BV. http://dx.doi.org/10.1016/j.jbi.2012.02.001. Disponível em: https://www.sciencedirect.com/science/article/pii/S1532046412000226. Acesso em: 6 nov. 2020.

SANTIAGO, M.A. **Guidelines e Protocolos Clínicos**. Anais do VIII Congresso Brasileiro de Informática em Saúde, Natal, 2008. Disponível em: http://www.avesta.com.br/tutorial/t10_1.pdf. Acesso em: 29 ago. 2020.

SANTOS, Andressa Maxwara Jovino dos; VECHIO, Gustavo Henrique del. INTELIGÊNCIA ARTIFICIAL, DEFINIÇÕES E APLICAÇÕES. **Revista Interface Tecnológica**, [S.L.], v. 17, n. 1, p. 129-139, 30 jul. 2020. Interface Tecnológica. http://dx.doi.org/10.31510/infa.v17i1.782. Disponível em: https://revista.fatectq.edu.br/index.php/interfacetecnologica/article/view/782. Acesso em: 16 maio 2021.

SAS. **Machine Learning**: o que é e qual sua importância?. O que é e qual sua importância?. 2021. SAS Insights. Insights sobre Análise de Dados. Disponível em: https://www.sas.com/pt_br/insights/analytics/machine-learning.html. Acesso em: 11 abr. 2020.

SAHAMI, Mehran *et al*. A Bayesian Approach to Filtering Junk E-Mail. In: AAAI WORKSHOP ON LEARNING FOR TEXT CATEGORIZATION, 1998, Madison. **Workshop.** Madison: AAAI Technical Report Ws-98, 1998. p. 55-62. Disponível em: https://www.aaai.org/Papers/Workshops/1998/WS-98-05/WS98-05-009.pdf. Acesso em: 15 out. 2020.

SCARDONI, Alessandro *et al*. Artificial intelligence-based tools to control healthcare associated infections: a systematic review of the literature. **Journal of Infection And Public Health**, [S.L.], v. 13, n. 8, p. 1061-1077, ago. 2020. Elsevier BV. http://dx.doi.org/10.1016/j.jiph.2020.06.006. Disponível em: https://www.sciencedirect.com/science/article/pii/S1876034120305001. Acesso em: 10 abr. 2020.

SILVA, Edna Lúcia da; MENEZES, Estera Muskat. **Metodologia da pesquisa e elaboração de dissertação.** 4. ed. Florianopolis: UFSC, 2005. 138 p.

SILVER SHARK SOLUTIONS. **Machine Learning**. 2018. Disponível em: https://silversharksolutions.com.br/index.php/solucoes-com-uso-machine-learning/. Acesso em: 21 ago. 2020.

SOUZA, Virginia Helena S. de; MOZACHI, Nelson. **O hospital**: manual do ambiente hospitalar. 3. ed. Curitiba: Manual Real Ltda, 2009. 874 p.

STONE, Peter *et al*. **Artificial Intelligence and life in 2030**: One hundred year study on Artificial Intelligence. Stanford Report of the 2015 Study Panel, California, USA, 52p., set. 2016. Disponível em: https://ai100.stanford.edu/2016-report. Acesso em: 24 mai. 2021

TAMAYO, Mario Tamayo y. **El processo de La investigación científica**: incluye evaluación y administración de proyectos de investigación. 4. ed. México: Limusa, 2003. 435 p.

THRUN, Sebastian *et al*. Stanley: the robot that won the Darpa grand challenge. **Journal Of Field Robotics**, [S.L.], v. 23, n. 9, p. 661-692, 2006. Wiley. http://dx.doi.org/10.1002/rob.20147. Disponível em: https://onlinelibrary.wiley.com/doi/abs/10.1002/rob.20147. Acesso em: 18 nov. 2020.

TUKEY, John W.. **Exploratory Data Analysis**. New Jersey: Addison-Wesley, 1977. 711 p.

UNIVERSIDADE FEDERAL DE SANTA. CATARINA. **Programa de Pós-Graduação em Tecnologias da Informação e Comunicação - PPGTIC:** Linhas de Pesquisa. 2021. Disponível em: http://ppgtic.ufsc.br/linhas-de-pesquisa/. Acesso em: 19 maio 2021.

URDAN, Timothy C.. **Statistics in plain English**. 3. ed. New York: Routledge, 2015. 224 p.

VAPNIK, Vladimir N.. The Nature of Statistical Learning Theory. **Springer-Verlag New York**, [S.L.], p. 1-314, 2000. Springer New York. http://dx.doi.org/10.1007/978-1-4757-3264-1.

WARNER, Jeremy L. *et al*. Classification of hospital acquired complications using temporal clinical information from a large electronic health record. **Journal Of Biomedical Informatics**, [S.L.], v. 59, p. 209-217, fev. 2016. Elsevier BV. http://dx.doi.org/10.1016/j.jbi.2015.12.008. Disponível em: https://www.sciencedirect.com/science/article/pii/S1532046415002889. Acesso em: 6 abr. 2020.

WIENS, Jenna; GUTTAG, John; HORVITZ, Eric. Patient Risk Stratification with Time-Varying Parameters: A Multitask Learning Approach. **Journal Of Machine Learning Research.** [S. L.], p. 1-23. abr. 2016. Disponível em: https://dl.acm.org/doi/pdf/10.5555/2946645.3007032. Acesso em: 25 maio 2021.

WHO, World Health Organization. **Worldwide country situation analysis**: response to antimicrobial resistance. Geneva: World Health Organization, 2015. 50 p. Disponível em: https://apps.who.int/iris/bitstream/handle/10665/163468/9789241564946_eng.pdf;jsessionid=C39E6493B082C8D37142345DB5DD6C79?sequence=1. Acesso em: 17 out. 2020.

APÊNDICE – A – CARTA DE ACEITE

Unimed Criciúma

Hospital Unimed Criciúma

Rua Estevão Emílio de Souza, 101
XXXXX-180 Bairro Ceará, Criciúma - SC
T. (48) 3478-3000
F. (48) 3478-3099

Criciúma, dezembro de 2019

CARTA DE ACEITE

Declaramos, para os devidos fins que se fizerem necessários, que concordamos em disponibilizar banco de dados, do Hospital Unimed Criciúma, localizado na Avenida Estevão Emilio de Souza nº 101 - Bairro Ceará, Criciúma/SC, para o desenvolvimento da pesquisa intitulada **"PROPOSTA DE UMA FERRAMENTA PARA AVALIAÇÃO DE RISCOS DE INFECÇÃO DO SÍTIO CIRÚRGICO UTILIZANDO TÉCNICAS DE INTELIGÊNCIA ARTIFICIAL"** sob a responsabilidade do pesquisador Jefferson Pacheco Dos Santos e supervisor Prof.ª Eliane Pozzebon do curso de Pós-Graduação em Tecnologias da Informação e da Universidade Federal de Santa Catarina, pelo período de execução previsto no referido projeto.

Daniela Loch Gomes

Enfª Daniela Loch Gomes
Coren-SC 249.075
Núcleo de Ética em Pesquisa
Hospital Unimed Criciúma

APÊNDICE – B – TABELA DE PROCEDIMENTOS

Procedimento	Total realizado	Total infecções	Taxa infecção
Colectomia	332	19	5.723
Microcirurgia - tumores intracranianos	61	3	4.918
Laparoscopia	352	15	4.261
Osteomielite	56	2	3.571
Gastrectomia	60	2	3.333
Fechamento de colostomia	134	4	2.985
Laparotomia	706	17	2.408
Enterectomia	87	2	2.299
Retossigmoidectomia	181	4	2.210
Artroplastia	582	11	1.890
Artrodese	331	5	1.511
Pieloplastia	67	1	1.493
Cateter duplo J	212	3	1.415
Hérnia de disco toraco-lombar	289	4	1.384
Esplenectomia	81	1	1.235
Gastrostomia	266	3	1.128
Hipospádia	183	2	1.093
Prostatavesiculectomia	193	2	1.036
Linfadenectomia	200	2	1.000
Prostatectomia	226	2	0.885
Nefrolitotripsia Percutânea	258	2	0.775
Artroscopia	288	2	0.694
Endometriose	370	2	0.541
Ressecção da próstata	765	4	0.523
Fratura	1149	6	0.522
Nefrectomia	210	1	0.476
Apendicectomia	1474	7	0.475
Herniorrafia	2499	9	0.360
Colecistectomia	3435	12	0.349

Gastroplastia	1519	5	0.329
Histerectomia	3052	10	0.328
Mastectomia	321	1	0.312
Abdominoplastia	1501	4	0.266
Prótese de mama	2970	7	0.236
Cesariana	12738	30	0.236
Ureterorrenolitotripsia	2767	6	0.217
Quadrantectomia	484	1	0.207
Cistoscopia	1545	3	0.194
Tireoidectomia	585	1	0.171
Reparo/reforço/sutura meniscal	1290	2	0.155
Exérese	3184	2	0.063
Histeroscopia	2946	1	0.034
Septoplastia	3738	1	0.027
Retirada de placas	148	0	0.000
Varizes	1538	0	0.000
Incontinência urinária	598	0	0.000
Refluxo gastroesofágico	970	0	0.000
Meniscectomia	56	0	0.000
Mastopexia	392	0	0.000
Lipoaspiração média	292	0	0.000
Acromioplastia	74	0	0.000
Hidrocele unilateral	292	0	0.000
Adenoamigdalectomia	1534	0	0.000
Vasectomia	1624	0	0.000

Fonte: Elaborado pelo AUTOR (2021).

ANEXO A - PROPOSTA DE CLASSIFICAÇÃO DA INTELIGÊNCIA ARTIFICIAL (IA) E APLICAÇÕES MEDIADAS POR IA NA MEDICINA E NA ATENÇÃO À SAÚDE DE ACORDO COM SEU CARÁTER BENÉFICO VERSOS PREJUDICIAL. SW = SOFTWARE, AR = REALIDADE AUMENTADA, VR = REALIDADE VIRTUAL, IOT = INTER.

AI and AI-mediated technologies	Specific implementations.	TAL	Social Impact
Algorithms for computer-aided diagnosis.	SW for decision support in (most) clinical areas.	8, 9	Positive
Structured reports, eHealth.	SW for improved workflow, efficiency.	8, 9	
AR/VR, advanced imaging tools.	Tools for information visualization and navigation.	6, 7, 9	
	Image-guided surgery. Teleoperation.	4, 6, 9	
Digital pathology, 'virtopsy'.	SW for automated, extensive analysis.	4-9	
Personalized, precision medicine.	Tailored treatments. Prediction of response.	4-9	
	'In-silico' modeling and testing. The 'digital twin'.	4-8	
	Drug design.	4, 8	
Apps, chatbots, dashboards, online platforms.	The 'digital doctor' (assistance for professionals and for patients).	8, 9	
Companion and social robots.	For hospitalized persons, children & the elderly.	4-9	
Big Data collection and analysis.	Epidemiology, prevention and monitoring of disease outbreaks.	2-9	
	Fraud detection. Quality control, monitoring of physicians and treatments.	4-9	
IoT, wearables, mHealth.	Automated clinical/health surveillance in any environment/institution.	7, 8	
	Monitoring, automated drug delivery.	7-9	
Gene editing.	Disease treatment, prevention.	7, 8	
Merging of medical and social data. 'Social' engineering.	Prevention of episodes with clinical relevance (e.g. suicide attempts).	6, 8	Controversial
	Tailored marketing (e.g. related to female cycles).	6, 8	
Reading and decoding brain signals. Interaction with neural processes.	Treatment of diseases. Restoring damaged functions.	3-8	
	Brain-machine inferfaces.	5-8	
	Control of prostheses, exoskeletons. 'Cyborgs'.	2-7	
	Neurostimulation. Neuromodulation.	4-8	
	Neuroprostheses (for the central nervous system).	2-5	
	Mind 'reading' and 'manipulation'.	1-3	
Genetic tests. Population screening.	Disease tests. Direct-to-consumer tests.	4-9	
Personalized, precision medicine.	Individual profiling. Personalized molecules (for treatment) at 'impossible' prices.	3-8	
Gene editing.	'Engineered' humans.	2, 6	
	Gene-enhanced 'superhumans'.	2	
	Self-experimentation medicine. Biohacking.	2, 6	
Fully autonomous AI systems.	The 'digital doctor'.	2-5	
	'Robotic surgeon'.	2, 4	
Human-animal embryos.	Organs for transplants.	2, 4, 5	
	Hybrid beings ('chimera').	2, 4	
The quest for immortality.	Whole-brain emulation / 'transplant'.	1, 2	
The search for artificial life forms.	'Living machines' ('biological robots', 'biobots')	4, 6	
	Military.	2, 3	
Evil biohacking.	Targeting specific individuals or groups.	1, 2	
Weaponization.	From 'small labs' to military labs.	1, 2	
Bioterrorism.	From 'small labs'.	1, 2	Negative

Fonte: GÓMEZ-GONZÁLEZ *et al.*, (2006, p.5).