UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ

MYLLER AUGUSTO SANTOS GOMES

PROPOSTA DE MODELO PREDITIVO PARA CUIDADOS DE SAÚDE BASEADO EM FUNCIONALIDADES DE ANÁLISE *BIG DATA* E TRANSFERÊNCIA DE TECNOLOGIA

PONTA GROSSA

2022

MYLLER AUGUSTO SANTOS GOMES

PROPOSTA DE MODELO PREDITIVO PARA CUIDADOS DE SAÚDE BASEADO EM FUNCIONALIDADES DE ANÁLISE *BIG DATA* E TRANSFERÊNCIA DE TECNOLOGIA

Predictive model based on big data analytics and technology transfer capabilities: an application for healthcare

Tese apresentada como requisito parcial à obtenção do título de Doutor em Engenharia de Produção, do Programa de Pós-Graduação em Engenharia de Produção, da Universidade Tecnológica Federal do Paraná.

Orientador: Prof. Dr. João Luiz Kovaleski

Coorientadora: Profa. Dra. Regina Negri Pagani

PONTA GROSSA

2022

Ministério da Educação
Universidade Tecnológica Federal do Paraná
Campus Ponta Grossa

MYLLER AUGUSTO SANTOS GOMES

PROPOSTA DE MODELO PREDITIVO PARA CUIDADOS DE SAÚDE BASEADO EM FUNCIONALIDADES DE ANÁLISE *BIG DATA* E TRANSFERÊNCIA DE TECNOLOGIA

Trabalho de pesquisa de doutorado apresentado como requisito à obtenção do título de Doutor Em Engenharia de Produção da Universidade Tecnológica Federal do Paraná (UTFPR). Área de concentração: Gestão Industrial.

Data de aprovação: 23 de setembro de 2022.

Dr. Joao Luiz Kovaleski, Doutorado - Universidade Tecnológica Federal do Paraná
Dr. Bruno Pedroso, Doutorado - Universidade Estadual de Ponta Grossa (UEPG)
Dr. Flavio Trojan, Doutorado - Universidade Tecnológica Federal do Paraná
Dr. Guilherme Moreira Caetano Pinto, Doutorado - Universidade Estadual de Ponta Grossa (UEPG)
Dra. Helyane Bronoski Borges, Doutorado - Universidade Tecnológica Federal do Paraná
Dra. Regina Negri Pagani, Doutorado - Universidade Tecnológica Federal do Paraná

Documento gerado pelo Sistema Acadêmico da UTFPR a partir dos dados da Ata de Defesa em 26/09/2022.

AGRADECIMENTOS

Primeiro a Ele, Deus, por me encorajar, motivar e alimentar o desejo de investigar condições de saúde, contribuindo para a nossa vida.

À minha amada esposa, Caroline, e nossos filhos lindos Louise (2 meses e meio) e Victor (9 anos). Vocês são a minha base para qualquer decisão e inspiração. Sem vocês não sei o que seria da minha vida. Obrigado por tudo.

Aos meus pais, Hamilton Cesar Gomes, *in memorian,* e Edilmair Santos, que são meus espelhos de dedicação, seriedade, comprometimento, formas de amor e meus modelos de vida.

Sou eternamente grato ao meu orientador, Prof. Dr. João Luiz Kovaleski, pelas contribuições; professor autêntico e inovador. Tenho uma admiração profunda pelo senhor.

À minha coorientadora, Profa. Dra. Regina Negri Pagani, pelas orientações e contribuições ao trabalho, pela sua disponibilidade, e que representa parte das minhas inspirações.

Aos membros do grupo de pesquisa GTT, pelos *feedbacks*, contribuições e reflexões em nossas reuniões. Estou desde 2015 com vocês, e o aprendizado só cresce. Muito obrigado.

Aos meus colegas de sala e aos colegas do PPGEP.

À secretaria do curso, pela cooperação e prontidão em todas as situações que precisei.

Aos professores do curso, com os quais muito aprendi; questiono-me, por qual motivo não comecei antes. Vocês me transformaram enquanto aluno, professor e pesquisador.

À Unicentro *Campus* Irati e aos colegas do Departamento de Administração, a quem agradeço pelo envolvimento e acolhimento por mais de oito anos na universidade. Muito, muito obrigado!

Meu agradecimento ao Hospital Santa Casa de Misericórdia de Ponta Grossa (PR), em especial aos setores de TI e Hospital de Ensino, por contribuir e permitir a realização desta pesquisa. Muito obrigado a todos!

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), por haver concedido, em meio a uma pandemia, uma bolsa (Código de Financiamento 001) para contribuir com o desenvolvimento da pesquisa.

“É somente nas misteriosas equações do amor que qualquer lógica ou razão pode ser encontrada. Você é a razão de eu estar aqui hoje, você é a razão de eu existir, você é todas as minhas razões.”

John Forbes Nash Jr.

RESUMO

Esta investigação tem como objetivo geral desenvolver um modelo preditivo apoiado pelas funcionalidades de análise *big data* e transferência de tecnologia (TT), visando desenvolver a capacidade de encontrar resultados sobre a ressubmissão clínica de pacientes diagnosticados com câncer e que foram submetidos aos cuidados de saúde. Para tanto, realizou-se uma pesquisa aplicada, orientada pela abordagem quantitativa, com caráter explicativo. O procedimento técnico é experimental e documental. Dados clínicos, dados de prontuários e registros hospitalares de câncer são subsídios necessários para o tratamento analítico por meio da aplicação de algoritmos de aprendizado de máquina, objetivando encontrar a relação entre variáveis, desenvolvendo a mensuração da taxa de ressubmissão clínica, associada a um processo de validação por meio de medidas de desempenho. Os resultados demonstram que o modelo preditivo possui boa precisão discriminativa. O algoritmo Regressão Logística Binária apresenta área sobre a curva AUC=0,890 para o conjunto de treinamento e no conjunto teste AUC=0,886. A árvore de decisão no conjunto de treinamento 75% apresenta índice Kappa 0.9992579, sensibilidade 0.7962441 e especificidade 0.191028; para o conjunto teste 25%, índice Kappa 0.9976999, sensibilidade 0.779661, especificidade 0.2. Conclui-se que o modelo preditivo possui poder de predição abrangente capaz de mensurar taxa de ressubmissão clínica, gerando informação significativa à gestão clínica e podendo reorientar o fluxo de trabalho do serviço de oncologia por meio de processos de TT.

Palavras-chave: análise *big data*; modelos preditivos; cuidados em saúde; transferência de tecnologia; aprendizado de máquina.

ABSTRACT

The general objective of this investigation is to develop a predictive model supported by big data analysis and technology transfer (TT) features in order to develop the ability to find results on the clinical resubmission of patients diagnosed with cancer who underwent health care. Therefore, an applied research was carried out, guided by the quantitative approach, with an explanatory character. The technical procedure is experimental and documentary. Clinical data, data from medical records and hospital cancer records are necessary subsidies for analytical treatment through the application of machine learning algorithms, aiming to find the relationship between variables by developing the measurement of the clinical resubmission rate associated with a validation process by through performance measures. The results demonstrate that the predictive model has good discriminative accuracy. The Binary Logistic Regression algorithm presents area under the curve AUC=0.890 for the training set and AUC=0.886 for the test set. The decision tree in the 75% training set has Kappa index 0.9992579, sensitivity 0.7962441 and specificity 0.191028, for the test set 25%, Kappa index 0.9976999, sensitivity 0.779661, specificity 0.2. It is concluded that the predictive model has comprehensive predictive power capable of measuring clinical resubmission rate, generating significant information for clinical management and being able to reorient the workflow of the oncology service through TT processes.

Keywords: big data analytics; prediction model; healthcare; technology transfer; machine learning.

LISTA DE FIGURAS

LISTA DE GRÁFICOS

LISTA DE QUADROS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

1.1 Contextualização e problema de pesquisa

Este tópico objetiva contextualizar a pesquisa, os elementos motivadores e determinantes para sua realização, fixação do objetivo geral e objetivos específicos, problema de pesquisa, magnitude do tema e justificativa para o seu desenvolvimento, bem como a aderência à área de engenharia de produção.

Na era digital, o grande volume de dados e informações que estão sendo gerado, em velocidades e variedades cada vez maiores, expressam a complexidade da análise *big data* enquanto prática organizacional (WANG; KUNG; BYRD, 2018). Desenvolvida a partir do conceito de inteligência de negócios e sistemas de suporte a decisão, a análise *big data* representa um conjunto de técnicas, tecnologias, sistemas, práticas, metodologias e aplicações com capacidade analítica para grandes conjuntos de dados em que procuram esclarecer o que está obscuro e subentendido para as organizações. Isto por meio da descoberta de informações sobre negócios, mercado e tomada de decisão (AMALINA *et al.,* 2019; MOHAMED *et al.,* 2020).

A análise *big data* e aprendizado de máquina desenvolvem grandes potenciais para serem utilizados de forma sistemática em conjuntos de dados, procurando descobrir padrões interessantes que até então eram desconhecidos. Além disso, descobrir informações nos armazenamentos de dados, objetivando melhorar a qualidade da saúde e redução de custos.

Uma alternativa para obter-se a capacidade tecnológica necessária à análise *big data* está no desenvolvimento de novas aplicações, como modelos preditivos, por exemplo, capazes de atender às expectativas das partes interessadas em situações nas quais os dados não sofrem o devido tratamento analítico (PANAYIDES *et al.,* 2018; RAZZAK; IMRAN; XU, 2020). Nesse sentido, as técnicas analíticas sofreram significativas mudanças. Atualmente, o aprendizado de máquina, uma abordagem computacional baseada na utilização de algoritmos capazes de fazer previsões e predições de resultados, são peças essenciais neste contexto (MCNUTT *et al.,* 2018).

Assim, os enfrentamentos da implementação de arquiteturas de análise *big data* e a remodelagem da capacidade tecnológica colocam organizações de saúde

em situação sensível para obtenção de benefícios, os quais transformam a prática clínica (GALETSI; KATSALIAKI; KUMAR, 2020).

Não somente demonstrar os resultados por meio da visualização de dados proeminentes da análise, mas também tornar acessível aos profissionais envolvidos no setor de saúde. A dependência da aquisição e interpretação dos dados, mesmo com as melhorias substanciais recebidas nos últimos anos, ainda sofrem dificuldades em virtude das características da análise *big data* incluindo volume, velocidade, variedade, variabilidade, veracidade, valor e valência (MOHAMED *et al*., 2020; SAGGI; JAIN, 2018).

Com essa complexidade desenhada pela diversidade de fonte de dados, novas técnicas são requeridas para lidar com os conjuntos de dados expressivos (WANG; HAJLI, 2017; SHAFQAT *et al.,* 2020).

A coexistência de técnicas analíticas avançadas está na combinação das linguagens de programação, como Python, Scala, R, SQL (JONES *et al.,* 2018; MOHAMED *et al.,* 2020).

Nesse contexto, a aplicação do algoritmo, desenvolvendo capacidade analítica apropriada à situação de investigação, irá produzir resultados os quais necessitam ser submetidos a medidas de desempenho para sua validação (WARING; LINDVALL; UMETON, 2020).

Contudo, com todas as suas funcionalidades estabelecidas, os processos de transferência de tecnologia aparecem como medida intrínseca ao resultado da análise (GUO; CHEN, 2019; PAN *et al.,* 2019; PASSOS *et al.,* 2019).

Profissionais de saúde necessitam de familiaridade com as técnicas computacionais, justamente para estruturar o processo de tomada de decisão e transferência de tecnologia (TT). É preciso avaliar a tomada de decisão algorítmica, devido à real necessidade de avaliação e interpretação sobre os resultados apresentados por parte dos profissionais envolvidos. Ainda assim, as medidas de TT precisam objetivar a significância clínica e a redução de custo voltadas ao tratamento do paciente, com monitoramento constante e melhoria de desempenho (MILLER; FRENCH, 2016; PAN *et al.,* 2019; HAQ *et al.,* 2020).

Diante das evidências descritas, esta pesquisa desenvolveu o seguinte problema de pesquisa: **Como a análise *big data* pode contribuir para a gestão e transferência de tecnologia, explorando grandes conjuntos de dados reais de**

pacientes acometidos por câncer, visando à predição de ressubmissão em nível clínico?

1.2 Objetivos da pesquisa

1.2.1 Objetivo geral

Desenvolver um modelo preditivo com funcionalidades de análise *big data* e transferência de tecnologia, direcionado para pacientes submetidos a cuidados de saúde.

1.2.2 Objetivos específicos:

- OE 1. determinar as principais características da análise *big data* e Transferência de Tecnologia (TT) no contexto de cuidados de saúde;
- OE 2. apresentar formas de aplicação de aprendizado de máquina, voltadas à análise de grandes conjuntos de dados;
- OE 3. construir *dataset* organizado pelos preditores estabelecidos com dados reais de pacientes submetidos a cuidados de saúde;
- OE 4. elaborar modelo preditivo com funcionalidades de análise *big data* e TT, apresentando o sistema de aprendizagem em saúde, orientado a análises preditivas;
- OE 5. analisar o desempenho do modelo preditivo, por meio de conjunto de dados reais de pacientes acometidos por câncer, com foco nas medidas de desempenho.

1.3 Justificativa da pesquisa

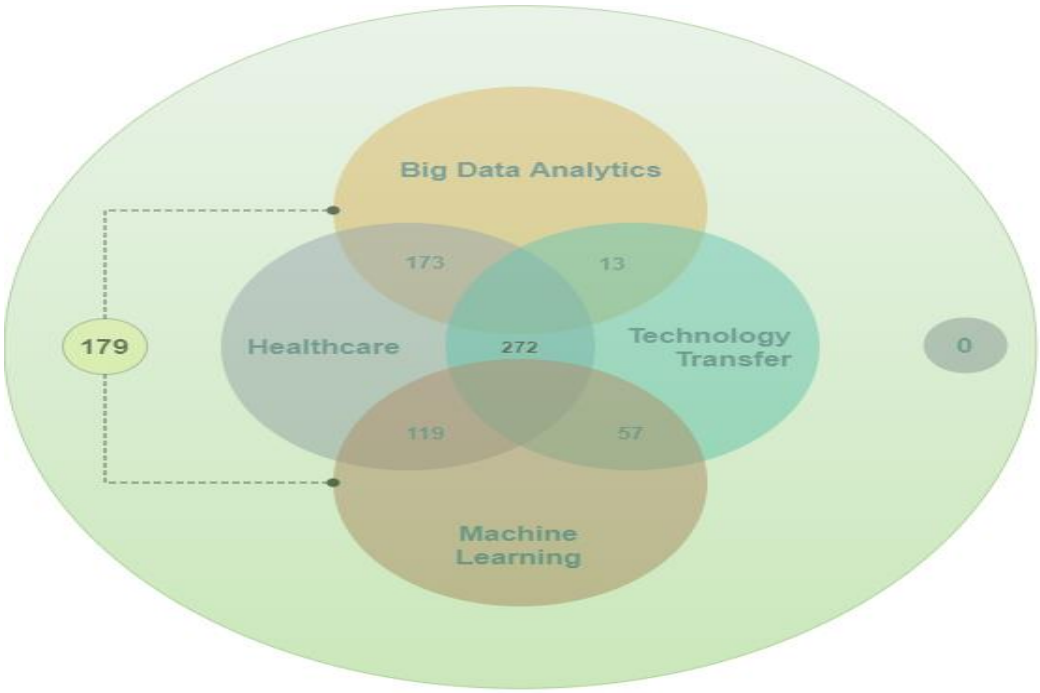
Esta pesquisa apresenta sua magnitude através da importância atribuída aos dados dos pacientes que precisam se submeter a alguma forma de tratamento analítico preditivo para saber se há previsão de retorno ou não de tais pacientes ao serviço de oncologia. Para isso, utilizaram-se de técnicas de análise *big data,*

transferência de tecnologia e aprendizado de máquina capazes de contribuir para a tomada de decisão, auxiliando a atuação médica no gerenciamento das equipes de enfermagem e multiprofissional. Nessa condição, um modelo preditivo com essas funcionalidades torna-se relevante, obedecendo orientações éticas vigentes.

O foco principal desta pesquisa é propor um modelo preditivo aplicado em resultados de pacientes submetidos a cuidados de saúde, com inter-relação de estágios, usufruindo do poder preditivo de algoritmos de aprendizado de máquina supervisionado. Com os desdobramentos da revisão sistemática de literatura, utilizando-se da metodologia *Methodi Ordinatio*, desenvolvida por Pagani, Kovaleski e Resende (2015), identificou-se um número pequeno de estudos sobre análise *big data* e TT em cuidados de saúde; a maioria dos estudos focaliza partes da análise *big data* aplicadas a determinadas doenças.

Investigou-se por meio da combinação das palavras-chave: "*big data analytics*", "*healthcare*", "*technology transfer*" e "*machine learning*" nas bases de dados *Science Direct*, *Scopus* e *Web of Science*, sendo que não foram localizados estudos que destacam totalmente os quatro contextos de investigação, conforme exposto na figura 1. Além disso, modelos preditivos semelhantes não foram encontrados.

Figura 1 - Originalidade da pesquisa



Fonte: Autoria própria (2022)

A partir dos resultados obtidos, percebeu-se a necessidade de desenvolver uma investigação, considerando os contextos de investigação, entendendo sua contribuição teórico-empírica. Foram coletados 800 trabalhos de forma sistemática, aplicando filtros orientados pela *Methodi Ordinatio.* Restaram 173 artigos com significativas contribuições para objetivos de pesquisa.

Para justificar o ineditismo e originalidade desta pesquisa, foram estabelecidas significativas características:

1. profunda investigação sobre transferência de tecnologia em cuidados de saúde;
2. proposta de desenvolvimento de um modelo preditivo composto de cinco estágios inter-relacionados (captura dos dados, processamento dos dados, análise dos dados, descoberta do conhecimento, tomada de decisão e transferência de tecnologia);
3. cada estágio é composto por algumas funcionalidades específicas, considerado um sistema de aprendizado em saúde para predição de ressubmissão clínica;

4. funcionalidades da análise *big data* e TT foram incorporadas ao modelo preditivo;
5. modelos preditivos raramente consideram a TT sobre os resultados das predições;
6. nova medida preditiva é constituída chamada de taxa de ressubmissão clínica;
7. modelo preditivo considera preditores de pacientes diagnosticados com câncer e que receberam primeiro tratamento, apurando probabilidades de retorno ou não ao serviço de oncologia.

Com o modelo preditivo desenvolvido, a intenção foi de contribuir com a atuação clínica em três linhas de interesse: do paciente, em obter informações precisas sobre seu tratamento e com possibilidade real de término; da equipe clínica, a qual possuirá ação supervisora com poder de orientação, baseado nas predições; do sistema de saúde, enquanto pagador, compreendendo a real aplicação dos recursos materiais e financeiros. Na engenharia de produção, as subáreas da engenharia organizacional e engenharia clínica norteiam o campo de investigação produzido por esta pesquisa, devido à sua preocupação com a gestão da informação em diferentes organizações; em especial, às voltadas aos cuidados de saúde e intensivas em dados; em produzir o devido tratamento analítico e contribuir para a tomada decisão e TT. O impacto social pode ser reconhecido pelos atores envolvidos.

1.4 Organização dos capítulos da tese

Este tópico da pesquisa possui como objetivo esclarecer as especificações estruturais da organização dos capítulos.

O capítulo 1 compreende a introdução, pergunta de pesquisa, objetivos de pesquisa e, de maneira finalística, justificativa à sua relevância.

O capítulo 2 compreende a revisão sistemática de literatura a qual corroborou com a pesquisa, e está estruturada da seguinte forma:

- transferência de tecnologia em organizações de cuidados em saúde;
- análise *big data*; e

- técnicas de aprendizado de máquina: conceitos e abordagens.

O capítulo 3 compreende os procedimentos metodológicos adotados nesta tese, e estão assim divididos:

- classificação da pesquisa;
- etapas da pesquisa; e
- procedimentos de coleta e análise dos dados.

O capítulo 4 estabelece os resultados encontrados e a análise e discussão para elaboração do modelo preditivo, dividido em:

- elaboração do *dataset* para construção do modelo preditivo;
- aplicação do modelo preditivo no *dataset bd_oncology*, possibilidades probabilísticas de algoritmos de aprendizagem supervisionada;
- discussão sobre o modelo preditivo proposto.

Por fim, o capítulo 5 destaca as considerações finais desta pesquisa, propondo possibilidades de estudos futuros.

2 REFERENCIAL TEÓRICO

Esta pesquisa apresenta um modelo preditivo para descobrir a probabilidade de um paciente retornar ou não ao serviço de oncologia e desenvolver TT. O modelo preditivo proposto é embasado nas seguintes bases teóricas: transferência de tecnologia em organizações de cuidados em saúde, características da análise *big data* e aprendizado de máquina. Sendo assim, nas seções subsequentes serão desenvolvidos esses tópicos.

2.1 Transferência de tecnologia em organizações de cuidados em saúde

Transferir tecnologia em contextos com especificações distintas exige, da organização e dos envolvidos, capacidade de percepção e utilidade voltadas à melhoria de desempenho a partir da força de trabalho. Permitindo, assim, fornecer diagnóstico, tratamento e acompanhamento preciso, alcançando os pacientes necessitados, com os avanços da tecnologia da informação e inteligência artificial. As inovações tecnológicas são cada vez mais presentes em organizações de cuidados em saúde (BIRKEN *et al.,* 2015).

Dessa maneira, o processo é capaz de desenvolver o compartilhamento de conhecimento, habilidades, tecnologias, *expertise*, métodos de fabricação e instalações entre universidades, empresas, governos e outras instituições, visando à garantia do desenvolvimento científico e tecnológico que caracteriza o fenômeno da transferência de tecnologia. Contudo nem todo participante é o candidato apropriado (JADHAV; GAUTAM; GAIROLA, 2014; BIRKEN *et al.,* 2015).

Ainda que os destinatários do objeto transferido devam possuir capacidades científicas e tecnológicas com intuito de facilitar o processo de transferência, rigorosos procedimentos de fabricação e treinamento, além de padrões regulatórios conhecidos, necessitam estar estabelecidos para garantir a eficácia e escalabilidade sobre o objeto transferido (CHOI *et al.,* 2015; BATTISTONI *et al.,* 2016).

Mover a tecnologia entre atores com interesses distintos torna-se um espaço de complexidade nas atividades de interação. Nesta lógica consistente, os sistemas nacionais de pesquisa têm sofrido, ao longo das últimas décadas, reformas com o intuito de aumentar a transferência de tecnologia e a comercialização da pesquisa

acadêmica para impulsionar os benefícios econômicos através de tecnologias inovadoras em saúde centrada na doença (ANCARANI *et al.,* 2016; NICOL *et al.,* 2016; DAVIES; RODERICK; HUXTABLE, 2019).

Não somente beneficiar um sistema de saúde ou arrecadar recursos financeiros para a organização, o enfoque da transferência de tecnologia em organizações de cuidados de saúde está na significância clínica descrita sobre o tratamento ao paciente, nos fatores de saúde e na eficácia de custos, buscando o melhor impacto de desempenho (MILLER; FRENCH, 2016). Do mesmo modo, a transferência de tecnologia de forma organizada é vista como impulsionador de organizações e pesquisadores a encontrar formas de comercialização de suas descobertas, permitindo o reconhecimento. Entretanto o compromisso da organização está no estabelecimento de melhores práticas, retenção de talentos e atratividade pela inovação (NILSEN *et al.,* 2016; HUANG *et al.,* 2018).

Com tendências emergentes, como *Internet* das coisas (*Iot – Internet of Things*), saúde inteligente, assistência médica inteligente, sensores não invasivos, dispositivos móveis, análise *big data* e robótica clínica, quando aplicadas, estabelecem funcionalidades as quais aumentam o nível de precisão e o custo aplicado a determinados procedimentos. Algumas funcionalidades apresentadas por Pan *et al.* (2019) são destacadas a seguir:

- fornecer supervisão operacional, gerenciamento médico e avaliação de desempenho;
- auxiliar no diagnóstico, percurso cirúrgico e compartilhamento de pesquisas para profissionais de saúde;
- fornecer monitoramento diário de saúde, gerenciamento de emergência e guia médico para pacientes em serviços diversificados.

No entanto, apesar de todas as vantagens oferecidas pelos serviços de saúde, os pontos de aceitação, utilidade percebida e risco percebido por profissionais de saúde representam o elemento transformador e cíclico acerca dos procedimentos convencionais de saúde. Isso aliado às estruturas tecnológicas disponíveis para utilização, à evolução para um modelo inteligente, apoiado pela tecnologia da informação e com decisões baseadas em evidências; ainda tem sua disseminação

nas organizações de cuidados de saúde de forma tardia (KAPOOR; LEE, 2013; PRIHODOVA; GUERIN; KERNOHAN, 2015; PAN *et al.,* 2019).

2.2 Análise *big data*

Com um exponencial movimento dos dados na *internet* e nas organizações, onde problemas de alta dimensionalidade estão disponíveis para análise e tomada de decisão, *big data* representa este universo em virtude de sua abundância, e tal qual uma amostra elevada se apresenta como um dos desafios técnicos para processar e extrair informações (CHAN; LU; WANG, 2018).

Alinhado a isso, o termo *big data analytics* se tornou comum no contexto da tecnologia da informação e na pesquisa acadêmica a partir de 2011 (DE MAURO *et al.,* 2018). Suas características essenciais, como informação, método, tecnologia e impacto são identificadas para representar seu significado geral em termos de desempenho e aplicação de análise prescritiva, descritiva e preditivas, visando obter intuições e projeções para orientar a tomada de decisão (ISTEPANIAN; AL-ANZI, 2018).

Por envolver diversas funções computacionais, a arquitetura de análise *big data* está associada à estrutura do ciclo de vida dos dados que se inicia com a captura dos dados; na sequência, pelo processamento representado pela transformação dos dados e, por fim, sua utilização, aqui chamada de consumo dos dados (MEHTA; PANDIT, 2018).

As indústrias pioneiras em computação de análise de *big data*, como bancos e comércio eletrônico, estavam começando a ter um impacto na melhoria dos processos de negócios e na eficácia da força de trabalho (KOZJEK *et al*., 2020), reduzindo os custos corporativos e atraindo novos clientes. No entanto a maior parte da criação de valor potencial ainda está em sua fase inicial, porque as técnicas preditivas de modelagem e simulação para analisar dados como um todo ainda não foram adequadamente desenvolvidas (KOZJEK *et al.,* 2020).

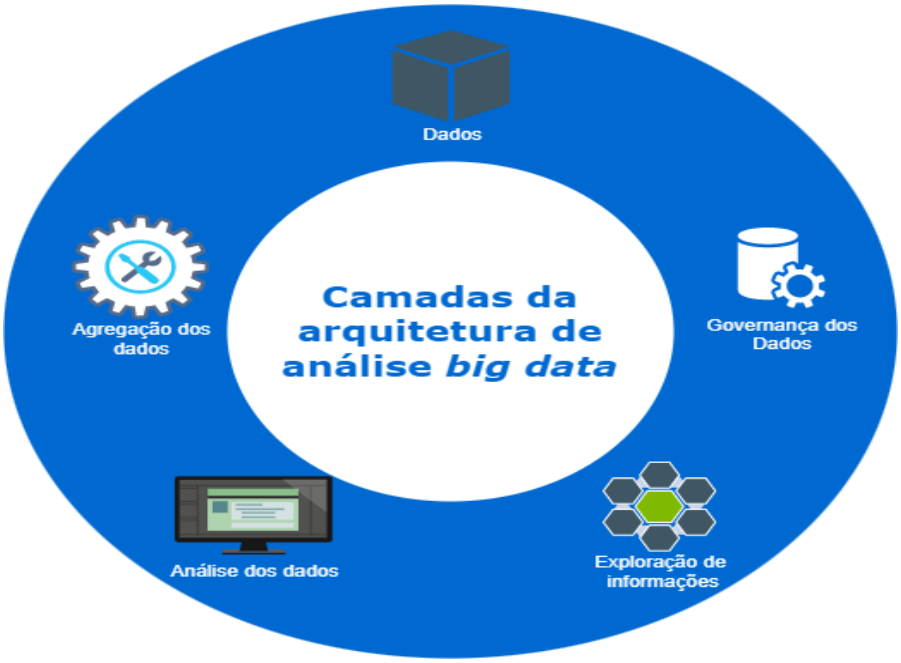
Por exemplo: em ambientes de saúde, os dados são capturados através de registros eletrônicos de saúde, imagens, medicamentos, procedimentos adotados. Para Rizwan *et al.* (2018), quando utilizados e armazenados em formatos heterogêneos, ou seja, formatos estruturados, não estruturados e semiestruturados,

discretos e contínuos, dados pessoais do paciente, diagnósticos, relatos clínicos, dados administrativos, prescrições, procedimentos, exames laboratoriais, imagens médicas, perfazem uma característica complexa e de grandes dimensões. Todavia a deficiência de dados nesses instrumentos pode representar um desafio às técnicas analíticas.

Essas condições dos dados exigem uma arquitetura capaz de preservar a sua essência e, ao mesmo tempo, ter a capacidade analítica para promover formas de processamento. Conceitualmente, uma base arquitetônica considera três componentes primordiais, sendo: agregação dos dados, análise dos dados e interpretação dos dados. Estes componentes são cobertos por funcionalidades específicas: a agregação de dados observa o processo de transformação para serem entendidos; na análise de dados, com a utilização de técnicas apropriadas, a formação de ideias é estabelecida e, por fim, a interpretação de dados que está associada à produção de relatórios e à visualização dos dados que subsidiarão, por sua vez, a tomada de decisão baseada em evidência (WANG *et al.,* 2018).

Um exemplo de arquitetura de análise *big data* é apresentado por Strang e Sun (2020) em que se consideram cinco camadas principais que, organizadas de forma lógica, permitem aos gestores a capacidade de entender a transformação dos dados e a consecução de fontes significativas de informações por meio do processo de análise *big data*, conforme apresentadas na figura 2:

Figura 2 - Camadas principais da arquitetura de análise *big data*



Fonte: Adaptado de Strang e Sun (2020)

Com capacidade de lidar com vários problemas e utilizando-se de inteligência computacional, a análise *big data* procura categorizar diferentes dados em diversas classes, identificando suas características e permitindo buscar inferências e significados em suas representações aqui identificadas como padrões. Dessa forma, as técnicas analíticas fornecem os *insights* necessários para produção de conhecimento sobre determinados fenômenos e suporte à tomada de decisão otimizada com criação de valor estratégico (STRANG; SUN, 2020).

A arquitetura de análise *big data* não é diferente de estruturas de tecnologia da informação (TI) convencionais. Porém há algo peculiar nas formas de processamento aplicadas: enquanto a estrutura de TI convencional envolve técnicas de *Business Intelligence* (BI) parciais ou completas, sua base está vinculada a um sistema autônomo. Quando o objeto de análise está inserido no *big data*, em virtude de seus atributos, necessariamente o processamento precisa desenvolver e executar o processamento distribuído (SHAFQAT *et al.,* 2020).

2.2.1 Características da análise *big data*

A relevância da matéria-prima utilizada na análise *big data* está associada à característica qualitativa e quantitativa de dados e informação, em virtude do aumento

exponencial do fluxo de informação com dados de variados tipos. Em sua maioria são dados não estruturados; consequentemente, o poder computacional em termos de armazenagem e processamento se tornou necessidade de uma sociedade baseada em informação e conhecimento orientados pela capacidade cognitiva humana (HADI *et al.,* 2020).

Os atributos do *big data* são incorporados à sua análise e considerados desafiadores no que tange à sua exploração analítica (AMALINA *et al.,* 2019; GALETSI; KATSALIAKI; KUMAR, 2019). Por essa razão, considerar volume, velocidade, valor, variabilidade, veracidade, variedade, pode determinar a técnica analítica apropriada. Outros autores, como Saggi e Jain (2018), apresentam valência como um atributo pertencente ao *big data,* destacando que, além dos 6Vs já apresentados, é necessário ter condições de conectividade aos dados, entendido como valência; surgindo, assim, 7Vs.

Volume: destaca a quantidade e imensidão dos dados produzidos em um determinado ambiente. Velocidade: faz menção ao movimento produzido pelos dados. Em termos estruturais das características dos dados, volume e velocidade assumem a variedade que apresentam; e, em termos qualitativos, a veracidade explora valores condizentes com a qualidade e, por fim, o valor, que necessariamente precisa apresentar coerência e entrega de valor para o ambiente, indivíduo ou organização (RISTEVSKI; CHEN, 2018; WARING; LINDVALL; UMETON, 2020).

Explorando a valência como atributo pertencente ao *big data*, impõem-se níveis de conectividade consistentes e intensificados a indivíduos e organizações, onde os quais apresentam tendências a fluxo de dados em tempo real (*streaming*). No entanto, para explorar este atributo, obstáculos computacionais, como rede, infraestrutura de TI, métodos de transformação dos dados precisam ser superados (SAGGI; JAIN, 2018; WANG; KUNG; BYRD, 2018).

Em virtude da complexidade do comportamento dos recursos, a tecnologia tem se tornado ampla e acessível, transformando-se em facilitador para o processo de surgimento do fenômeno *big data*. Tecnologias *open-source* surgem com a intenção de baratear a computação distribuída e a disponibilidade de dados. Assim, o surgimento da arquitetura *Hadoop*, uma estrutura de *cluster* de máquinas chamada de computação paralela, coopera para o desempenho analítico (MOHAMED *et al.,* 2020).

Observando os métodos aplicados, estes envolvem a adoção de técnicas analíticas que permitem a transformação de grandes conjuntos de dados e informações, possibilitando a identificação de padrões que permitem uma interpretação passível de ser transferida (GHANI *et al.,* 2019). Dentre os métodos recentes utilizados, encontramos a análise de *clusters*, algoritmos, processamento em linguagem natural, reconhecimento facial, visual, sonoro, redes neurais, modelagem preditiva, modelos de regressão, análise de rede social, análise de sentimento, processamento de sinal e visualização de dados (CHRIMES; ZAMANI, 2017).

Movimentos dos atributos do *big data*, das ferramentas computacionais e dos métodos analíticos aplicados representam as características da análise *big data.* Uma organização apresentada por Ghani *et al.* (2019) destaca formas de como os métodos analíticos podem ser classificados:

- análise descritiva, objetivando esclarecer o que já aconteceu;
- análise de diagnóstico, procura entender os acontecimentos;
- análise preditiva, compreende tendências e o que pode acontecer futuramente;
- análise prescritiva, pesquisa melhores resultados, dada certa condição.

Desse modo, observar o surgimento do aumento informacional, do poder tecnológico e computacional disponível, e dos avanços dos métodos analíticos que norteiam a análise *big data* em diversos ramos da ciência moderna, como economia, saúde, finanças, produção, dentre outros. Esses ramos contribuem para a construção da arquitetura deseja com base nas descobertas almejadas (SAGGI, JAIN, 2018; HADI *et al.,* 2020).

Além disso, existe um apelo às aplicações de aprendizado de máquina, a qual procura aprender com os dados, utilizando técnicas estatísticas e algoritmos capazes de analisar conforme a caracterização do problema. Este apelo é relacionado à exploração analítica (ROTH *et al.,* 2018). Em termos de precisão dos dados, a preservação do estado em que se encontram antes da transformação necessária para sua análise, contribui para a técnica de aprendizado de máquina aplicada, permitindo desenvolver a compreensão e exames detalhados (AJAYI *et al.,* 2020).

2.2.2 Geração e aquisição de dados

Os avanços que a tecnologia da informação desenvolveu nas últimas décadas permitiu considerar não somente a infraestrutura de *hardware* e *software,* mas também a observância dos dados, o que se tornou significativamente relevante devido à sua complexidade e contribuição para a economia baseada em dados (IP *et al.,* 2018).

Com níveis de conectividade elevados, associados ao uso intensivo da TI, indivíduos e organizações se tornaram motores de geração de dados em diversos lugares. Exemplo disso são as plataformas de mídias sociais, dados sensores de máquinas, dispositivos eletrônicos, dados biométricos, dados gerados por humanos, dentre outros. Essa diversidade de fontes estabelece a natureza em que os dados se apresentam, destacando os formatos estruturado, semiestruturado e não estruturado (GALETSI; KATSALIAKI; KUMAR, 2020; SHAFQAT *et al.,* 2020).

Considerando a natureza dos dados, estes, no formato estruturado, podem advir das possibilidades de geração citadas acima, com definições claras e precisas. Por exemplo: dados empresariais, dados de sistemas gerenciais, como gestão de relacionamento com clientes, sistemas de gestão empresarial, sistema de gestão da cadeia de suprimentos, entre outros, os quais possuem representações estabelecidas (SAHEB; IZADI, 2019).

No formato não estruturado, não existem padronizações estabelecidas, mas sim uma variedade significativa de formatos os quais exigem um esforço computacional considerável: vídeos, textos, sensores, mídias sociais são alguns exemplos (NAIR; SHETTY; SHETTY, 2018; MOHAMED *et al.,* 2020).

Dados semiestruturados necessitam de técnicas não convencionais. O estabelecimento de regras gerais, apoiadas por técnicas de inteligência artificial, é necessário para desenvolver a dinâmica do processo de análise. Exemplos de dados semiestruturados: dados de transações comerciais, transação eletrônica de documentos (EDI), XML (*Xtensible Markup Language*), JSON etc., conforme se pode visualizar através da figura 3.

Figura 3 - Natureza dos dados e exemplos

Fonte: Adaptado de Mohamed *et al.* (2020)

No entanto, dadas as possibilidades de geração de dados citadas acima, o fator comum entre elas está associado à confiança da conectividade com a rede de transmissão de dados; fato este que permite elevar a qualidade dos serviços prestados e a sua personalização (HADI *et al.,* 2020; KWON; LEE; KIM, 2017).

A necessidade do armazenamento correto, expansivo e confiável entre os diversos formatos dos dados exige a adoção de banco de dados específicos capazes de atender à escalabilidade do *big data*. Essa condição se apresenta como um desafio tecnológico que exige investimentos em sistemas de gerenciamento de dados e *know-how* específico (TABESH; MOUSAVIDIN; HASANI, 2019).

Estruturas de armazenamentos de dados são fundamentais para gerenciar o *big data*, justamente para aplicações corretas dos métodos analíticos. Ainda assim, os dados crescem exponencialmente e, portanto, estruturas convencionais de armazenamento e gerenciamento de dados não são compatíveis. Devido a isso, estruturas e sistemas de banco de dados, como NowSQL, MongoDB, HBase, Cassandra, HDFS (*Hadoop Distributed File System*) foram desenvolvidos

(HARERIMANA *et al.,* 2018). Porém a solução com maturidade, escalonável e viavelmente financeira está na computação em nuvem (YACCHIREMA *et al.,* 2018).

Normalmente, objeções de dimensionalidade e interoperabilidade dos dados são existentes e ocorrem devido à variedade de formatos. A dimensionalidade é um problema-chave para análise *big data*, pois funções de frequência, totalidade, exploração e dependência estão intrinsecamente associadas (HARERIMANA *et al.,* 2018). Sendo assim, a frequência refere-se a quantidades de vezes em que os dados são analisados pelo usuário, objetivando melhor precisão dos algoritmos aplicados; totalidade é relacionada ao desejo do usuário de processar e analisar todos os dados disponíveis; aplicação de métodos analíticos pelo usuário representa a exploração, e, por fim, equilibrar investimentos entre metodologias existentes, e novas metodologias podem desenvolver a função de dependência ao usuário (SUBUDHI; ROUT; GHOSH, 2019).

Superada a dimensionalidade, a interoperabilidade está ligada à capacidade de integração dos dados. E tal integração, em análise, é necessária (HARERIMANA *et al.,* 2018; JENA, 2020). Alguns tipos de interoperabilidade são destacados pela revisão sistemática de literatura, observados pelo estudo de Jena (2020), sendo:

- funcional – quando se tem nomenclaturas ou interpretações diferentes;
- metadados – em um banco de dados relacional à coluna, possui um nome que é semelhante aos metadados de conteúdo da coluna;
- instância – acrônimos e códigos não possuem o mesmo significado.

Objeções de interoperabilidade e seus tipos, normalmente ocorrem em países em desenvolvimento, em virtude das diferentes estruturas computacionais utilizadas com nomenclaturas distintas, dificultando o processo de análise e se tornando inviável em um fluxo de dados em tempo real (*streaming*).

2.2.3 Processamento e análise de dados

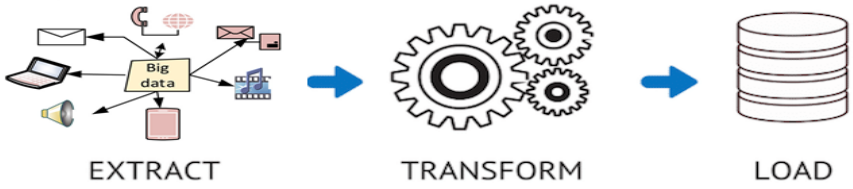
Processar e analisar dados em grandes proporções exige níveis computacionais elevados para obter a veracidade das aplicações tecnológicas corretas. Devido à variedade acentuada dos dados estruturados, semiestruturados e

não estruturados, processá-los para obter confiabilidade da análise é mais que necessário (BABAR *et al.,* 2018).

Desse modo, os dados precisam ser processados de maneira que permita o desenvolvimento da exploração com a tendência de buscar significados relativos ao objetivo da análise. Isto se apresenta como um desafio, e esse processo de mineração inicial é necessário para consolidar os dados que, em um primeiro momento, se apresentam como impuros e, se analisados, podem estabelecer incertezas na análise (MCNUTT *et al.,* 2018). Para o funcionamento adequado das técnicas analíticas, o processamento de dados está associado à capacidade de integração, sendo esta dotada de funcionalidades, como aquisição, transformação e armazenamento para que sua consistência seja preservada ao longo do tempo (WANG *et al.,* 2018).

Essas funcionalidades-chave são descritas como ferramentas de extração-transformação-processamento (ETL) e se comportam como *middleware,* uma aplicação que atua entre o sistema operacional e os *softwares* com fins de apoiar a estabilidade necessária às aplicações de análise, conforme figura 4 (TABESH; MOUSAVIDIN; HASANI, 2019; SHAFQAT *et al.,* 2020; SIVAPARTHIPAN *et al.,* 2020). Esse papel de limpeza é fundamental para preparação do grande conjunto de dados, antes da aplicação de técnicas analíticas. Os algoritmos, por exemplo. Sendo assim, a maior dificuldade está na insuficiência analítica em dados no setor de saúde, contribuindo para problemas e causando danos às medidas de prevenção de doenças (RAZZAK; IMRAN; XU, 2020).

Figura 4 - Processo de extração, transformação e processamento (ETL)



Fonte: Adaptado de Sivaparthipan *et al.* (2020)

Já há consenso entre os cientistas que o processamento de dados leva mais tempo que a implementação do modelo de análise, em virtude do alto nível de ausências e inconsistências (RIZWAN *et al.,* 2018; DUBEY *et al.,* 2020). As principais fontes de dados em cuidados de saúde são os registros médicos, os exames laboratoriais e as imagens com relatórios os quais podem conter valores ausentes.

Por exemplo: dados não fornecidos pelos pacientes, negligência de dados por parte do hospital, erros de máquina e erros humanos (WANG *et al.,* 2018).

Uma abordagem comum utilizada é ignorar dados ausentes e focalizar nos casos completos e disponíveis (CHEN *et al.,* 2017), e outro método citado é a imputação de valor único, em que os valores perdidos são substituídos por valor alternativo, como zero, média e moda (HARERIMANA *et al.,* 2018).

Em termos de avanços significativos dos métodos utilizados e disponíveis, a abordagem imputação baseada no modelo de análise *big data* apresenta um método de predição que permite atingir os valores próximos aos possíveis dados reais com a ajuda de dados disponíveis (RIZMAN *et al.,* 2018). Outros métodos de imputação são: o modelo de fator latente, baseado em características do paciente, e o método de imputação múltipla, usando equações encadeadas (RIZMAN *et al.,* 2018; JOHNSTON *et al.,* 2019).

A significância do processamento de dados consegue estabelecer o método de análise de dados a partir de capacidades analíticas que indivíduos e organizações possuem. Esse aspecto contribui para aplicação de ferramentas, técnicas e procedimentos que são voltados ao processamento, organização, armazenamento, visualização e análise de dados com o objetivo de fornecer *insights* que subsidiam a tomada de decisão e instrumentos de planejamento em todos os níveis estratégicos (WANG *et al.,* 2018; JOHNSTON *et al.,* 2019; SIVAPARTHIPAN *et al.,* 2020).

Analisar dados é esclarecer um universo obscuro e subentendido, através do processamento computacional. Com a maior disponibilidade de fontes de dados, se tornou interessante aplicar técnicas analíticas capazes de contribuir com a melhoria da gestão de recursos, qualidade do atendimento e precisão na tomada de decisão (ALLAREDDY *et al.,* 2019). Entender o comportamento de recursos, processos e pessoas está na observação e análise dos dados produzidos. Nessa ótica, havendo um número significativo de fontes de dados, maior será a variedade e a precisão dos dados, assegurando o que geralmente é denominado *big data* (BABAR *et al.,* 2018).

A onipresença do *big data* exige da análise *big data* uma alternativa analítica capaz de promover escalabilidade em virtude da natureza dos dados; e a computação paralela em uma plataforma pode ser a solução (HARERIMANA *et al.,* 2018).

Ainda assim, alguns estudos classificaram em quatro grupos de pesquisa sobre *big data*, colocando a análise de dados em grande escala como primeiro grupo

de pesquisa. Para esclarecer suas funcionalidades e diversas técnicas analíticas utilizadas, Zhang *et al.* (2019) destaca as seguintes:

1. aprendizado de máquina – máquina de suporte de vetor (SMV), análise de *clusters*, algoritmos de regressão e classificação, redes neurais;
2. *data warehouse* e metadados – *Extensible Markup Language* (XML), *web mining*;
3. mineração de dados em grande escala – processamento do fluxo de dados, visualização dos dados, análise de série temporal e mineração de texto.

As classificações apresentadas não esgotam as possibilidades sobre as técnicas analíticas aplicadas à análise de dados em grande escala. No entanto esta apresentação inicial coloca a perspectiva da transferência de conhecimento e tecnologia, possibilitando, entre os diferentes interessados em um ecossistema de informação complexo, a orientação por significados dos conteúdos os quais permitem transformar processos e subsidiar uma tomada de decisão baseada em evidências (HUANG *et al.,* 2018; GOMES *et al.,* 2019; PAN *et al.,* 2019).

2.3 Técnicas de aprendizado de máquina: conceitos e abordagens

Considerada uma subárea da Inteligência Artificial (IA), o campo do conhecimento do aprendizado de máquina tem promovido soluções para investigar grandes conjuntos de dados, colocando resultados esperados produzidos por máquinas através algoritmos, em vez de decisões subjetivas e arbitrárias criadas por seres humanos (MA; ZHANG; WANG, 2014).

Em termos conceituais, Kibria *et al.* (2018) esclarecem que aprendizado de máquina é colocar máquinas na condição de aprendizado próprio, a partir de iniciativas que provocam tentativas e erros. Para os autores mencionados acima, o poder de extração de informações significativas com a imputação de dados brutos revela verdadeiras preciosidades pela funcionalidade de correlações de várias fontes de dados.

Uma forma típica de aprendizado de máquina está na estruturação de fases, sendo a obtenção de dados a primeira fase, seguida das atividades de extração de

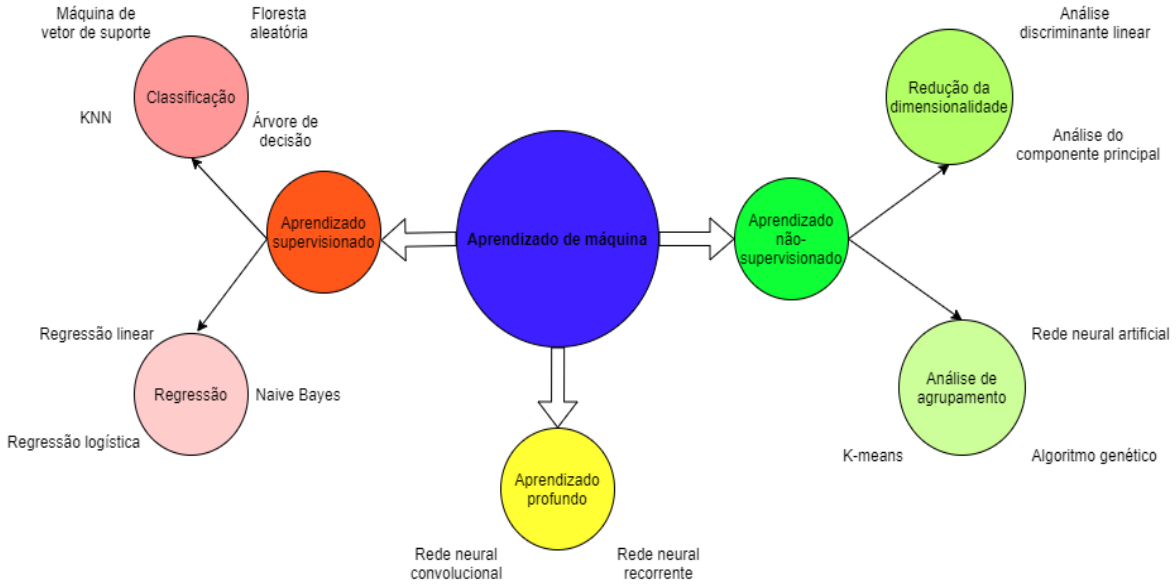
dados com preceitos de utilização na entrada de um modelo de aprendizado de máquina representando a segunda fase. Após isto, o modelo é treinado, efetuado testes e otimizado com base nas avaliações aplicadas sobre os resultados e na experiência do especialista, que de alguma forma consome um tempo incerto destacando a terceira fase (LIU *et al.,* 2017; YU; WANG; LIEW, 2019).

Outra ótica conceitual destaca que aprendizado de máquina é um campo que estuda métodos para buscar e explorar os padrões encontrados em um conjunto de dados. Padrões estes que são utilizados para expandir a compreensão do mundo real a partir do comportamento dos dados; e tal compreensão apoiada por áreas como ciência da computação, estatística e otimização, tendo em sua essência uma relação entre problemas de otimização e conjunto de dados (WIENS; SHENOY, 2018).

Expandindo a compreensão conceitual, Jones *et al.* (2018) argumenta que o aprendizado de máquina se preocupa com algoritmos aplicados em computadores, estabelecendo o aprendizado entre os relacionamentos ou padrões complexos descobertos ou não, sustentado por uma base de dados e procurando tomar decisões com precisão. Outra forma está na manifestação do processo geral de aprendizado de máquina, que se inicia com o estabelecimento de dados e rótulos, separando conjunto de dados em duas possibilidades: treinamento e teste. O treinamento representa o espaço para que o processo de aprender e modelar seja desenvolvido, baseado em algoritmo, enquanto o teste procura realizar as avaliações de desempenho para consolidação dos resultados finais (KIBRIA *et al.,* 2018; AHMED *et al.,* 2020; HIDALGO *et al.,* 2020).

De maneira unânime, o aprendizado de máquina é classificado em três abordagens de aprendizagem: supervisionada, não supervisionada e profunda, conforme o grau de intervenção humana necessária e formas de rotulação dos dados para o cumprimento de atividades. Nessa ponderação, essas abordagens são enquadradas em grupos os quais consideram a forma de supervisão e rotulação sobre o aprendizado produzido (fig. 5). A classificação e a regressão são consideradas aprendizagem supervisionada, enquanto técnicas de agrupamento e redução da dimensionalidade e análise de agrupamento são baseadas em aprendizagem não supervisionada (AKHAVIAN; BEHZADAN, 2015; KOSE; GOKTURK; KILIC, 2015; JONES *et al.,* 2018; AHMED *et al.,* 2020; LEITE *et al.,* 2020).

Figura 5 - Uma estrutura de aprendizado de máquina, métodos e algoritmos



Fonte: Adaptado de Gao *et al.* (2020)

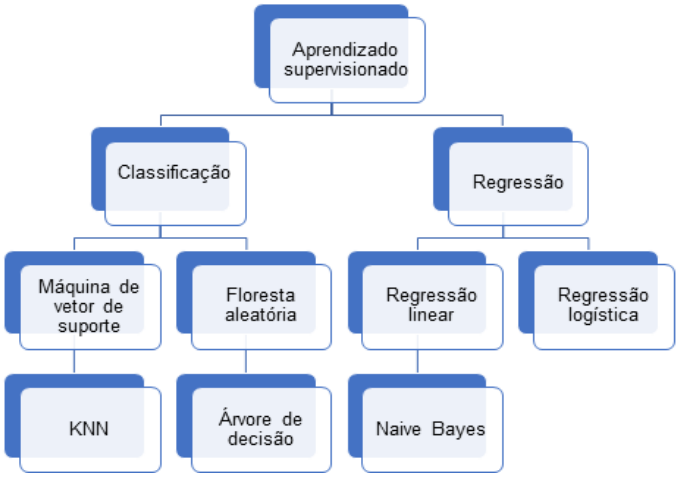
Expandindo os grupos de aprendizado de máquina, avanços consideráveis foram desenvolvidos em diversas investigações com a intenção de observar e analisar os dados que não estavam cobertos por técnicas convencionais. Por exemplo: aprendizado semissupervisionado que utiliza dados não rotulados para compreender a distribuição de probabilidade no espaço de entrada do modelo (KIBRIA *et al.,* 2018; LEE; SHIN; REALFF, 2018; SHARMA *et al.,* 2020). O aprendizado por reforço preocupa-se com a relação ótima para um objetivo construído entre os dados de entrada e suas parcialidades de resposta sobre os rótulos. Destaca-se a funcionalidade de simultaneidade entre as fases de aprendizagem, em que se caracteriza um recurso de auto-otimização em um ambiente incerto e dinâmico, observando as respostas em tempo real e ações ótimas e não ótimas que são tomadas. Após isso, aprende de forma implícita e explícita, associando o custo em um determinado estado do sistema (BERGER *et al.,* 2020; KIBRIA *et al.,* 2018; LEE; SHIN; REALFF, 2018; SHARMA *et al.,* 2020).

A análise *big data* possui relação intrínseca com o aprendizado de máquina, e isto se dá pela capacidade analítica que os algoritmos podem desenvolver em um grande conjunto de dados através de ferramentas computacionais avançadas. Desse modo, considerar uma estrutura complexa de dados e, em alguns casos subutilizada, pode fornecer soluções que transformarão a prática clínica através da descoberta de novos conhecimentos práticos.

2.3.1 Algoritmos para abordagem supervisionada

Com avanços consideráveis nas diversas áreas da computação nas últimas décadas, os algoritmos passaram por um processo de evolução em que foram sendo adequados às tarefas e problemas, baseados na sua capacidade de aprendizagem em termos de classificação e regressão (fig. 6). No entanto o que move a capacidade analítica é a quantidade de dados utilizados no treinamento (FERNÁNDEZ *et al.,* 2020).

Figura 6 - Relação entre os métodos e principais algoritmos de aprendizagem supervisionada



Fonte: Adaptado de Gao *et al.* (2020)

O que se torna determinante na escolha do algoritmo e a forma como o problema se apresenta são as possibilidades do algoritmo aplicado nos cuidados em saúde, orientados por abordagem de aprendizagem. Diante disso, o delimitante sobre os algoritmos utilizados foi a revisão sistemática de literatura, atividade que permitiu analisar a aplicação de aprendizado de máquina, conforme apresentado no quadro 1 e, em seguida, a apresentação dos conceitos e funcionalidades.

Quadro 1 - Algoritmos classificados por abordagem de aprendizagem supervisionada

Tipo	Supervisionado	Autor
Algoritmo	Regressão logística **Regressão linear**	**Abdelaziz *et al.*** (**2018**) Angelillo *et al.* (2019) Hatton *et al.* (2019) Kwakernaak *et al.* (2020) Hadi *et al.* (2020) **Gasimova e Abbasli** (**2020)** Johnston *et al.* (2019) Khan *et al.* (2018) Li *et al.* (2020) Passos *et al.* (2019)
	Naive Bayes	Alotaibi *et al.* (2020) Hadi *et al.* (2020) Gardner e Padman (2020) Gasimova e Abbasli (2020) Khan *et al.* (2018) Junior *et al.* (2019) Li *et al.* (2020) Suthaharan (2020)
	Árvore de decisão	Yavaraj e Sripreethaa (2019) Gasimova e Abbasli, (2020) Haq *et al.* (2020) Hidalgo *et al.* (2020) Khan *et al.* (2018) Li *et al.* (2020) Moyo *et al.* (2018) Chen *et al.* (2020)
	Floresta aleatória	Angelillo *et al.* (2019) Bdzok e Loannidis, (2019) Lima e Delen, (2020) Haq *et al.* (2020) Mazumdar *et al.* (2020). Kashi *et al.* (2020) Mcwillians *et al.* (2019)
	Máquina de vetor de suporte	Angelillo *et al.* (2019) Li *et al.* (2019) Li *et al.* (2020) Souri *et al.* (2020) Sujitha e Seenivasagam (2020) Gasimova e Abbasli, (2020) Srivastav; Singh; Suri (2019) Venkatesan, Srinivasan e Rajendiran (2019).
	KNN	Angelillo *et al.* (2019) Ali *et al.* (2020) Moreira *et al.* (2019) Gasinova e Abbasli (2020) Li *et al.* (2020) Zubar e Balamurugan (2020)

Fonte: Autoria própria (2022) (grifo do autor)

Levando-se em consideração essa classificação de algoritmos, todos foram observados nas investigações encontradas através da revisão sistemática de literatura. A regressão logística é considerada uma abordagem estatística convencional, mas com enorme capacidade preditiva. Sua atuação é voltada a modelar a relação entre variáveis, e suas previsões se tornam precisas quando o resultado rotulado (piora do paciente em nível clínico) é construído por uma combinação linear dos fatores (ANGELILLO *et al.,* 2019; JOHNSTON *et al.,* 2019; KWAKERNAAK *et al.,* 2020).

O modelo logístico é baseado na função logística. Para Hadi *et al.* (2020), esta função é zero (0) quando x - ∞. Entretanto, quando a função é um (1), $x + \infty$ a função logística é estabelecida pela seguinte equação:

$$f(x) = \frac{1}{1+e-x} \quad (1)$$

Para estimar a probabilidade em problemas com rótulos, o intervalo do modelo logístico torna-se adequado devido ao intervalo produzido (HADI *et al.,* 2020; GASIMOVA; ABBASLI, 2020). O índice de recursos combinados é x onde é dado como soma linear:

$$x = \beta 0 + \beta 1F1 + \beta 2F2 + \beta 3F3... + \beta NFN \quad (2)$$

Portanto β0 expressa y interceptar β1...βη estabelecendo os coeficientes de regressão $f1...fn$ para relatar as variáveis de recursos e n é o número total de recursos no modelo de previsão (KHAN *et al.,* 2018; HADI *et al.,* 2020; JOHNSTON *et al.,* 2019; GASIMOVA; ABBASLI, 2020). A probabilidade condicional pode ser descrita como:

$$P(C - c \mid Fi = fi) = \frac{1}{1e-(\beta 0+\sum_{i=1}^{n} \beta ifi)} \quad (3)$$

Hadi *et al.* (2020) esclarecem que a probabilidade condicional no qual $P\,(C{=}c \mid Fi{=}fi)$ expressa a probabilidade condicional de uma determinada variável de classe $C = c$ dado vetor de característica FV, com isso se $C = 1$ logo a probabilidade condicional para $C = 0$ sendo:

$$P(C = 0 \mid Fi = fi) = 1 - P(C = 1 \mid Fi = fi) \quad (4)$$

Regressão linear é um algoritmo de aprendizado de máquina, considerado comum, sendo amplamente utilizado. Seu objetivo é estabelecer uma relação entre as variáveis de forma linear, mas, infelizmente, este algoritmo captura somente um tipo de relação, promovendo uma dependência linear e baixa capacidade de processamento (ABDELAZIZ *et al.,* 2018; GASIMOVA; ABBASLI, 2020).

Conhecido por ser baseado no teorema de Bayes, algoritmo *Naive Bayes* é um classificador estatístico probabilístico que atua com uma série de características independentes obtidas em um conjunto de dados históricos voltado a determinar a probabilidade de uma situação ou incidente (HADI *et al.,* 2020; ALOTAIBI *et al.,* 2020; SUTHAHARAN, 2020). Sobre a denominação ingênuo, se justifica devido a assumir as variáveis de características que não estão relacionadas entre si (ALOITABI *et al.,* 2020; HADI *et al.,* 2020).

Focado na probabilidade posterior e anterior, sua notação matemática é estabelecida:

$$P(C = c \mid Fi = fi) = P(C = c)\prod_{i=1}^{n} P(Fi = fi \mid C = c) \quad (5)$$

Para probabilidade anterior, quando $P(C = c)$ representa uma situação anterior, e a probabilidade de F dado C ocorre em:

$$P(Fi = fi|C = c) = \frac{\sum_{i=1}^{n}(C=c \wedge Fi=fi)}{\sum_{i=1}^{n}(Ci=Ci)} \quad (6)$$

A representação da probabilidade conjunta se dá pela expressão:

$$\prod_{i=1}^{n} P(Fi = fi \mid C = c) \quad (7)$$

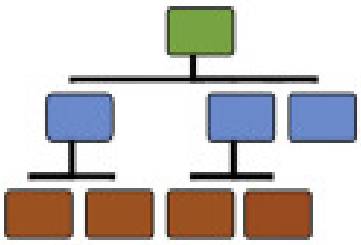
As vantagens deste algoritmo estão na baixa variância, na aplicação computacional rápida, implementação simples e funciona de forma adequada quando

situações apresentam um número elevado de dimensões. *Naive Bayes* pressupõe que a presença de uma característica na classe não está vinculada à presença de outra característica (ALOTAIBI *et al.,* 2020; GASIMOVA; ABBASLI, 2020).

Semelhante à estrutura de uma árvore com vários galhos entrelaçados e perfazendo um fluxograma, o algoritmo árvore de decisão denota em cada nó interno ou nó não folha um teste ou atributo. Cada ramo é usado para apresentar o resultado sobre teste ou atributo e, por fim, nó de folha contém o rótulo de classe que indica o valor final previsto. Esse algoritmo pode ser aplicado a grandes conjuntos de dados de forma eficiente (YUVARAJ; SRIPREETHAA, 2019; HAQ *et al.,* 2020). Quando está em construção, a árvore de decisão atua em conjuntos de treinamento e teste, e ambos são utilizados para estabelecer a delimitação do tamanho ideal necessário. Sendo assim, um dos atributos é selecionado para configurar um nó-raiz para servir de base para a construção da árvore (MOYO *et al.,* 2018; YUVARAJ; SRIPREETHAA, 2019; HIDALGO *et al.,* 2020).

Esse algoritmo pode utilizar variáveis numéricas e categóricas e atuar com valores ausentes na etapa de treinamento. Os relacionamentos não lineares entre as variáveis garantem o desempenho da árvore, estabelecendo as relações de forma intuitiva e de fácil compreensão devido à sua representação gráfica explicativa, conforme figura 7 (HIDALGO *et al.,* 2020).

Figura 7 - Configuração da árvore de decisão



Fonte: Adaptado de Hidalgo *et al.* (2020)

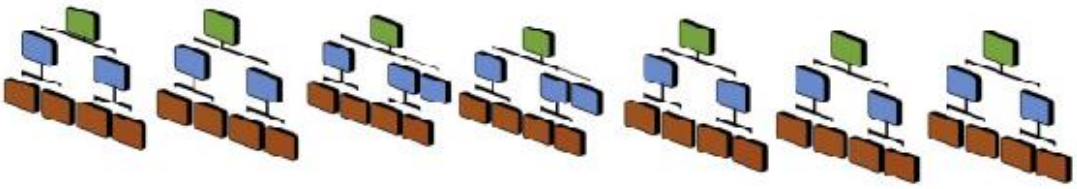
Outro algoritmo encontrado nas investigações da revisão sistemática de literatura com abordagem supervisionada é a floresta aleatória. Com poder de inferência e previsão, este algoritmo é articulado por modelo de conjuntos o qual é composto por uma variedade de pequenas árvores capazes de desenvolver resultados testáveis e robustos. Geralmente é reconhecido pelas comparações com

outros algoritmos em termos de desempenho superior (BZDOK; IOANNIDIS, 2019; KEMPA-LIEHR *et al.,* 2020).

Com funcionalidades aderentes a grandes conjuntos de dados, dados imperfeitos, constantemente preocupado com o nível de previsão, a floresta aleatória fornece a compreensão da mecânica por trás dos modelos preditivos (LIMA; DELEN, 2020). Existe uma dependência acerca do poder preditivo das árvores de decisão, ocasionando uma tendência de generalização. Para entender seu funcionamento, a cada nó dentro da floresta, as variáveis de entrada são escolhidas aleatoriamente, sendo que os nós são divididos conforme os critérios de impureza interna durante o crescimento das árvores da floresta. Os resultados são produzidos através de cada árvore, fornecendo um voto de classificação. Com isso, os votos de todas as árvores são agrupados e, assim, a classe que possuir o máximo de votos é apurada (ANGELILLO *et al.,* 2019; KASHI *et al.,* 2020; LIMA; DELEN, 2020; MAZUMDAR *et al.,* 2020).

Para demonstrar a sua representação gráfica, a figura 8 é apresentada:

Figura 8 - Floresta aleatória e seus conjuntos de árvores



Fonte: Adaptado de Lima e Delen (2020)

Máquina de Vetores de Suporte (MVS) é frequentemente utilizada para fazer previsões. Sua atuação inicia-se com a formação de um hiperplano em um espaço n-dimensional que normalmente representa uma linha que divide as variáveis conforme suas classes. Em especial, este espaço n-dimensional classifica de forma significativa os pontos de dados. Seu objetivo é encontrar o melhor limite entre os dados e calcular a distância possível entre todas as categorias. Não é sensível a outros pontos de dados (VENKATESAN; SRINIVASAN; RAJENDIRAN, 2019; GASIMOVA; ABBASLI, 2020; SOURI *et al.,* 2020).

Sua forma de classificação é atraente devido à sua capacidade de lidar com pequenos conjuntos de dados de treinamento, e sua flexibilidade, em termos de modelar sistemas complexos, é considerada uma técnica de aprendizado de máquina não paramétrica, por causa de o número de parâmetros se tornar crescente com a

quantidade de dados de treinamento (LI *et al.,* 2019; SOURI *et al.,* 2020). Para esclarecer a função de previsão utilizada, Li *et al.* (2019) explicam que previsões são baseadas em algumas funções no espaço de entrada, e o aprendizado está no processo de inferir os parâmetros dessa função (demonstrada abaixo):

$$y(x) = \sum_{n}^{N} = 1 \; wn \; K(x, xn) + \epsilon \tag{8}$$

Compreendendo a função acima, wn são os pesos do modelo que ligam o espaço de recursos à saída, k é a função kernel e ϵ é o termo de ruído independente, bastante utilizado em práticas de diagnóstico e prognóstico na saúde. Sua utilização se dá como ferramenta de regressão para valores contínuos, os quais são reconhecidos como regressão de vetor de suporte (ANGELILLO *et al.,* 2019; LI *et al.,* 2019; LI *et al.,* 2020).

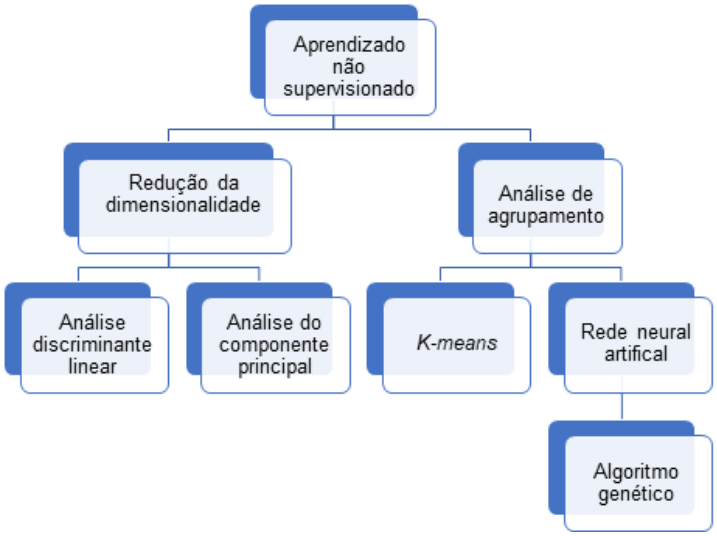
Baseado em analogias para desenvolver o aprendizado, KNN chamado de *K-Nearest-Neighbors* é um algoritmo classificador, em que o conjunto de treinamento consiste em vetores n-dimensionais. Para determinar a classe de um novo caso que não pertence ao conjunto de treinamento, o KNN busca k elementos do conjunto de treinamento que estejam mais próximos desse elemento. Assim, com menor distância (MOREIRA *et al.,* 2019; ALI *et al.,* 2020). Para fins de esclarecimentos, k elementos são conhecidos nos NNs. Na sequência, é verificada as classes desses k vizinhos, e a classe com maior frequência é atribuída à classe do novo caso. Este processo de classificação pode ser computacionalmente limitado se considerar um grande conjunto de dados; mas para outras aplicações, este algoritmo é bem aceitável (MOREIRA *et al.,* 2019; ALI *et al.,* 2020; GASIMOVA; ABBASLI, 2020; LI *et al.,* 2020; ZUBAR; BALAMURUGAN, 2020). A equação apresentada destaca o cálculo da distância euclidiana para pontos de dados.

$$d\,(x, y) = \sqrt{(x1 - y1)^2 + w2(x2 - y2)^2 + \ldots + (xn - yn)^2} \tag{9}$$

2.3.2 Algoritmos para abordagem não supervisionada

Algoritmos de aprendizagem não supervisionada são caracterizados por não estabelecerem rótulos ou resultados prévios para fins de processamento. Para isso, encontrar padrões e extrair a estrutura oculta dos dados sem qualquer rotulagem especializada se tornam o objetivo principal (ROTH *et al.,* 2018). A figura 9 apresenta os métodos pertencentes à aprendizagem não supervisionada e seus principais algoritmos.

Figura 9 - Relação entre os métodos e principais algoritmos de aprendizagem não supervisionada e seus métodos



Fonte: Adaptado de Gao *et al.* (2020)

Associado a isso, o quadro 2, a seguir, demonstra alguns algoritmos de aprendizagem não supervisionada encontrados na revisão sistemática de literatura.

Quadro 2 - Algoritmos classificados por abordagem de aprendizagem não supervisionada

Tipo	Não supervisionado	Autor
Algoritmo	Análise do componente principal	Roth *et al.* (2018) Savin *et al.* (2018)
	Análise discriminante linear	Tharwart *et al.* (2017)
	K-means	Lancaster *et al.* (2019) Fernández *et al.* (2020) Sharma *et al.* (2020)
	Rede neural artificial	Chen *et al.* (2017) Bem-Assuli *et al.* (2019) Fan *et al.* (2019) Li *et al.* (2019) Sun *et al.* (2019) Qaisar *et al.* (2020)

	Algoritmo genético	AHMED *et al.* (2020)

Fonte: Autoria própria (2022)

A redução da dimensionalidade expressa um conjunto de algoritmos de extração de características. E tal conjunto procura eliminar subconjuntos de atributos em espaços de alta dimensão. Por exemplo: o espaço bidimensional e tridimensional, com tendências à clusterização de dados. Existe um número significativo de dimensões em grandes conjuntos de dados, o que exige das técnicas de redução das dimensões para não degradação do desempenho do algoritmo selecionado (ROTH *et al.,* 2018; SAVIN *et al.,* 2018).

Dentre as principais técnicas de redução da dimensionalidade, destaca-se a Análise do Componente Principal (ACP). Trata-se de um algoritmo de transformação linear que procura produzir novos recursos, compreendidos como componentes principais, estabelecendo a sua variância máxima dos dados. Sua funcionalidade é orientada pela matriz de transformação que mapeia o espaço d-dimensional e projeta um novo espaço k-dimensional; outros categorizados em linear, como ACP; análise discriminante linear, análise de componente independente, análise fatorial, não linear. Isso dividido em dois grupos: global como escala multidimensional e local como análise discriminante de *Kernel Fish*, sendo uma extensão da análise discriminante linear (ROTH *et al.,* 2018; ANOWAR; SADAOUI; SELIM, 2021).

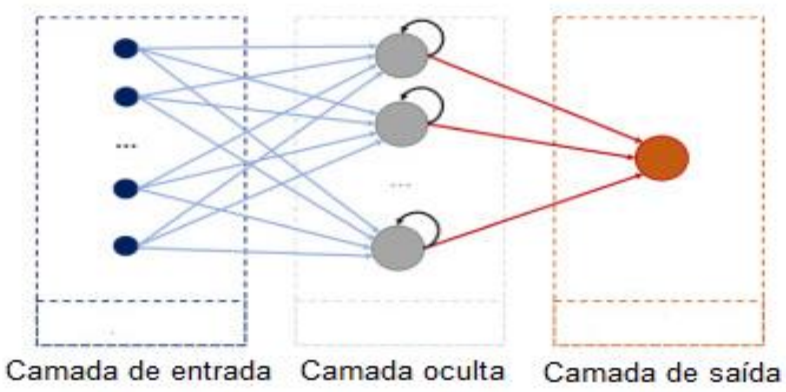
Na maioria das vezes, a Análise Discriminante Linear (ADL) é definida como linear. Sua atuação é voltada para a identificação do novo espaço de recursos e para projetar dados com objetivo de maximizar a separabilidade das classes, conhecida como recursos dependentes. O mesmo ocorre para os recursos independentes. Eles extraem novos recursos independentes que mais separam classes e, assim, desenvolve duas matrizes de dispersão: primeiro para as classes, onde calcula a distância média de cada classe; e a segunda para ser inserida dentro da classe, onde retorna a distância média de cada classe e os dados dessa classe (SAVIN *et al.,* 2018; ANOWAR; SADAOUI; SELIM, 2021).

O algoritmo *K-means* é considerado um método clássico de agrupamento, justamente pela sua simplicidade e eficácia em seu desenvolvimento através de linguagem de programação (SHARMA *et al.,* 2020). Conhecido como um algoritmo voltado para análises de agrupamento com comportamento iterativo, o *K-means* atua com escolha de forma aleatória de k objetos como centro de agrupamento inicial. Após

esta etapa, o algoritmo calcula a distância entre cada objeto e cada centro de agrupamento, atribuindo cada objeto ao centro de agrupamento mais próximo. A representação do *cluster* ocorre através dos centros de agrupamento e dos objetos atribuídos (LANCASTER *et al.,* 2019; GAO *et al.,* 2020).

Uma das formas de vincular as unidades de entrada às unidades de saída é uma característica marcante da Rede Neural Artificial (RNA). Com camadas interconectadas (fig. 10), tal rede expõe uma série de arestas e nós ponderados. Todos esses possuem pesos em cada *link* e, justamente nesse espaço, o aprendizado ocorre através do ajuste dos pesos (BEN-ASSULI *et al.,* 2019; QAISAR; SUBASI, 2020).

Figura 10 - Representação gráfica da rede neural artificial



Fonte: Adaptado de Li *et al.* (2019)

Com a finalidade de representar de forma matemática a atividade do cérebro humano, a RNA considera neurônios artificiais como unidades de processamento, interconectados pelas camadas de entrada. Ainda assim, articula os dados pré-processados e atua como *link* para as camadas ocultas. Estas, compostas por neurônios que possuem um modelo matemático para determinar sua saída com base nas entradas, podendo expressar combinações lineares ponderadas pelos pesos, utilizando uma função de ativação. Neste caso, a função sigmoide é utilizada para preditar resultados positivos sobre os dados pré-processados ativando linearmente a camada de saída (BEN-ASSULI *et al.,* 2019; LI *et al.,* 2019; SUN *et al.,* 2019; QAISAR; SUBASI, 2020).

Considerado um algoritmo de abordagem de aprendizagem não supervisionada, o algoritmo genético foi inspirado na teoria da evolução natural de Charles Darwin, e é voltado a solucionar problemas de padronização e otimização em

dados restritos e irrestritos, transformando iterativamente uma população e encontrando soluções individuais (AHMED *et al.,* 2020). Em termos de regras, esse algoritmo se baseia na pesquisa heurística, com três regras: seleção de elementos de dados (pai), regras de cruzamento para os dados e mutação para mudança aleatória (MEHTA; PANDIT; SHUKLA, 2019; AHMED *et al.,* 2020).

No entendimento dos autores citados acima, o algoritmo genético constantemente é utilizado para problemas de classificação e com capacidade de otimização e seleção de recursos. Ainda assim, possui alguns desafios que lhe impõem limitações computacionais, como sistema de pontuação não convencional e dificuldades no conjunto de dados de treinamento de classificadores. No entanto a sua aplicação em diversas áreas da medicina tem sido amplamente considerada.

2.3.3 Algoritmos para abordagem de aprendizagem profunda

Para esclarecer este conjunto de algoritmos de abordagem de aprendizagem profunda, conforme observado na revisão sistemática de literatura, os agrupamentos de algoritmos ocorreram pela forma de aprendizagem aplicada através do algoritmo utilizado. Contudo, nesta pesquisa, consideraram-se os algoritmos encontrados na revisão sistemática de literatura, ou seja, os encontrados são pertencentes à aprendizagem profunda, conhecida como *Deep Learning*.

Quadro 3 - Algoritmos de aprendizagem profunda

Tipo	Profundo	Autor
Algoritmo	Rede neural convolucional	Brunetti *et al.* (2019) Jenna (2020) Liu *et al.* (2017) Padarian, Minasny e Mcbratney (2019)
	Rede neural recorrente	Lee, Shin e Realff (2018) Jena (2020) Leite *et al.* (2020)

Fonte: Autoria própria (2022)

Redes neurais convolucionais são uma arquitetura de aprendizado profundo e são capazes de classificar imagens em diferentes categorias, com base em seu comportamento difuso; aprendendo, sim, de forma automática através de camadas convolucionais, sendo um espaço o qual combina múltiplos processos não lineares. Recursos computacionais são intensamente utilizados em termos de custos e tempo,

sendo que existem duas abordagens consideradas pela literatura: a) transferência de aprendizado "*transfer learning*", em que o retreinamento é aplicado em modelos pré-treinados em diferentes categorias; b) extratores de características, que são constituídos de várias camadas convolucionais as quais criam diferentes camadas de representações de recursos. Neste espaço, a captura das saídas é utilizada em outros algoritmos se tornando novas possibilidades de resultados (LIU *et al.,* 2017; PADARIAN; MINASNY; MCBRATNEY, 2019).

Normalmente, a rede neural convolucional procura obter dados na forma de várias matrizes. Por exemplo: imagens. A camada convolucional executa as convoluções sobre uma matriz, utilizando vários filtros, e é interconectada por pesos, sendo aprendidos durante o treinamento. Após o treinamento, os filtros podem identificar características diferentes e semelhantes (PADARIAN; MINASNY; MCBRATNEY, 2019).

Aplicável também ao universo de palavras, Jena (2020) destaca que se xi é o vetor de palavras em uma dimensão k correspondente à $iésima$ palavra em um texto; um texto com $'n'$ palavras foi representado como $x1: n = x1 \oplus x2 \oplus ... xn$ sendo $\oplus$ um operador de concatenação. Para unificar uma representação matricial de textos em diferentes comprimentos; um comprimento máximo de todos os textos foi usado como comprimento fixo às matrizes de texto. Em textos curtos, um vetor zero foi preenchido na parte de trás de uma matriz de texto. Após isto, a operação de convolução em cada matriz de texto é usada para transformá-la em uma escala ci, onde sua notação é estabelecida da seguinte forma:

$$xci = f(w.xi : i + h - 1 + b)$$

$$c = [c1,\ c2...cn - h + 1] \quad (10)$$

Para esclarecer as variáveis inseridas na notação matemática acima, Jena (2020) explica que a variável w é um mapa de filtro, $'h'$ evidencia o tamanho da janela de um filtro, $'f'$ é uma função de ativação não linear e $'b'$ representa um termo de polarização. A partir disto, os vetores regionais de palavras $xi: i + h$ - 1 na matriz de texto foram convoluídos para $c1$. A subamostragem foi assim norteada:

$$cmax = \{c\} \quad (11)$$

Com a equação acima, um filtro gera um *cmax* de uma matriz de texto, e as operações de convolução e subamostragem são realizadas (JENA, 2020).

Redes neurais recorrentes de uma arquitetura de aprendizado profundo, com a sua estrutura subjacente às redes neurais artificiais atuam com dados de entrada sequenciais, como fala e linguagem; não precisam ser independentes um do outro. Um aprimoramento da precisão ocorre por meio da descoberta do conhecimento adquirido nas iterações anteriores (LEE; SHIN; REALFF, 2018). Este tipo de algoritmo utiliza o conhecimento obtido na última análise para realizar a mesma tarefa para cada elemento de uma sequência, direcionando a memória para captura de informações processadas.

Uma descrição de Jena (2020) esclarece que o processamento ocorre através de um estado oculto *'h'* e uma saída opcional *'y'* que opera em uma sequência de comprimento variável $x = (x1, \ldots xN, \ldots xT)$. A cada passo de cada vez *'t'* o estado oculto $h(t)$ da rede neural recorrente é atualizado por $h(t) = f(h\,(t\text{-}1), xEU)$, quando a função f é uma função de ativação não linear. Exemplos: a Estrutura da Memória de Longo Prazo (LSTM) e a Unidade Recorrente Fechada (GRU). Após ter inserido palavra vetor no modelo, a distribuição de polaridade da polaridade global pode ser dada pela camada *softmax* usando $h(T)$ através da seguinte notação:

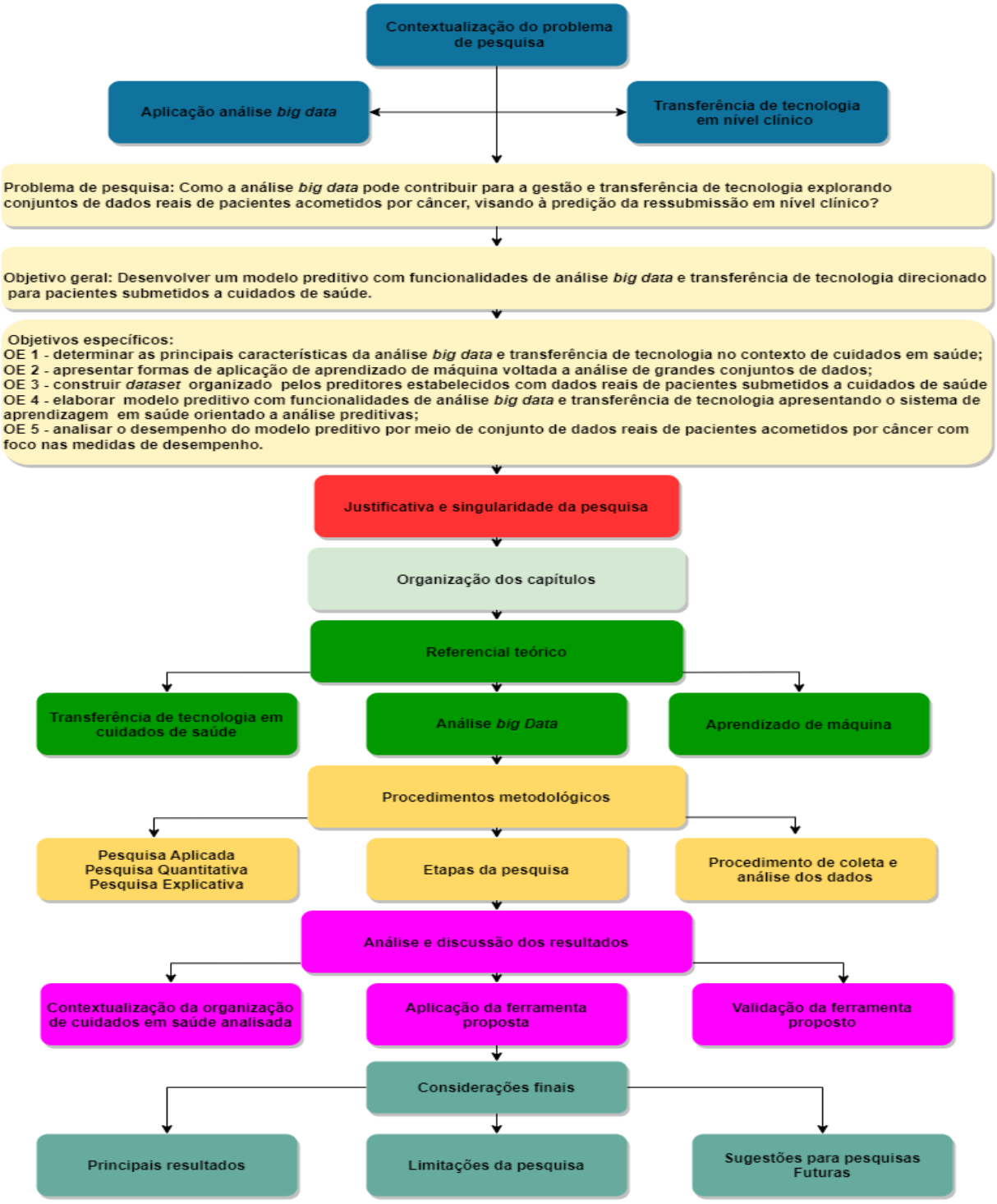
$$p(yi = 1)\frac{exp(wih(T))}{\sum_{j}^{k} exp(wJh(T))} \quad (12)$$

k é o número de classes, $J=o,\ldots\,(K\text{-}1)\ e\ (K\text{-}1)$ e wJ são matrizes de peso *'W'* da camada *softmax;* quando reutiliza as unidades ocultas das camadas anteriores, este algoritmo permite codificar informações passadas, além de possibilitar comprimir uma entrada de comprimento variável em um vetor de comprimento fixo $h\,(T)$.

3 METODOLOGIA

A disposição dos elementos da pesquisa é apresentada na figura 11, elencando todas as fases de desenvolvimento.

Figura 11 - Disposição dos elementos da pesquisa



Fonte: Autoria própria (2022)

3.1 Classificação da pesquisa

Sobre a ótica para abordar a natureza da pesquisa, a classificação **aplicada** é apropriada devido ao seu desenvolvimento em uma determinada realidade, objetivando contribuir com processos e conhecimentos produzidos em uma situação real (GIL, 2019).

Na abordagem do problema, a classificação é **quantitativa.** Para Creswell (2010), essa abordagem se define pela aplicação dos pontos fortes quantitativos, em observar a amostra e generalizar para uma população por meio de experimentos, possibilitando maior compreensão sobre o problema estudado.

Acerca dos objetivos, esta pesquisa é classificada como **explicativa**, por esclarecer e identificar elementos que aproximam ao problema de estudo, conhecendo de forma detalhada as variáveis do estudo, explorando métodos e técnicas aplicadas em um fenômeno da realidade, porque explica a razão das situações (GIL, 2019).

Quanto aos procedimentos técnicos, a pesquisa é estabelecida como **experimental e documental,** devida à aplicação de técnicas estatísticas visando à predição da piora em nível clínico através de uma ferramenta, além de contribuir com a compreensão do que ocorre ou por que ocorre a utilização de dados de um conjunto de pacientes em determinado processo de tratamento. Geralmente, esses dados estão inseridos em bancos de dados e não receberam nenhum tipo de tratamento analítico. Para a pesquisa experimental, Gil (1999) esclarece que o delineamento está na determinação de um objeto de estudo, na seleção das variáveis que seriam capazes de influenciá-lo, estabelecendo formas de controle e de observação dos efeitos que a variável produz. Aliada a isto, a utilização de técnicas estatísticas é predominantemente aceita.

A pesquisa documental possui duas estratégias de coleta de dados: o local onde os documentos são encontrados e coletados, possibilitando dois ambientes, o campo ou o laboratório; e a segunda estratégia refere-se à fonte dos dados, podendo estar nos documentos ou no campo (APPOLINÁRIO, 2009).

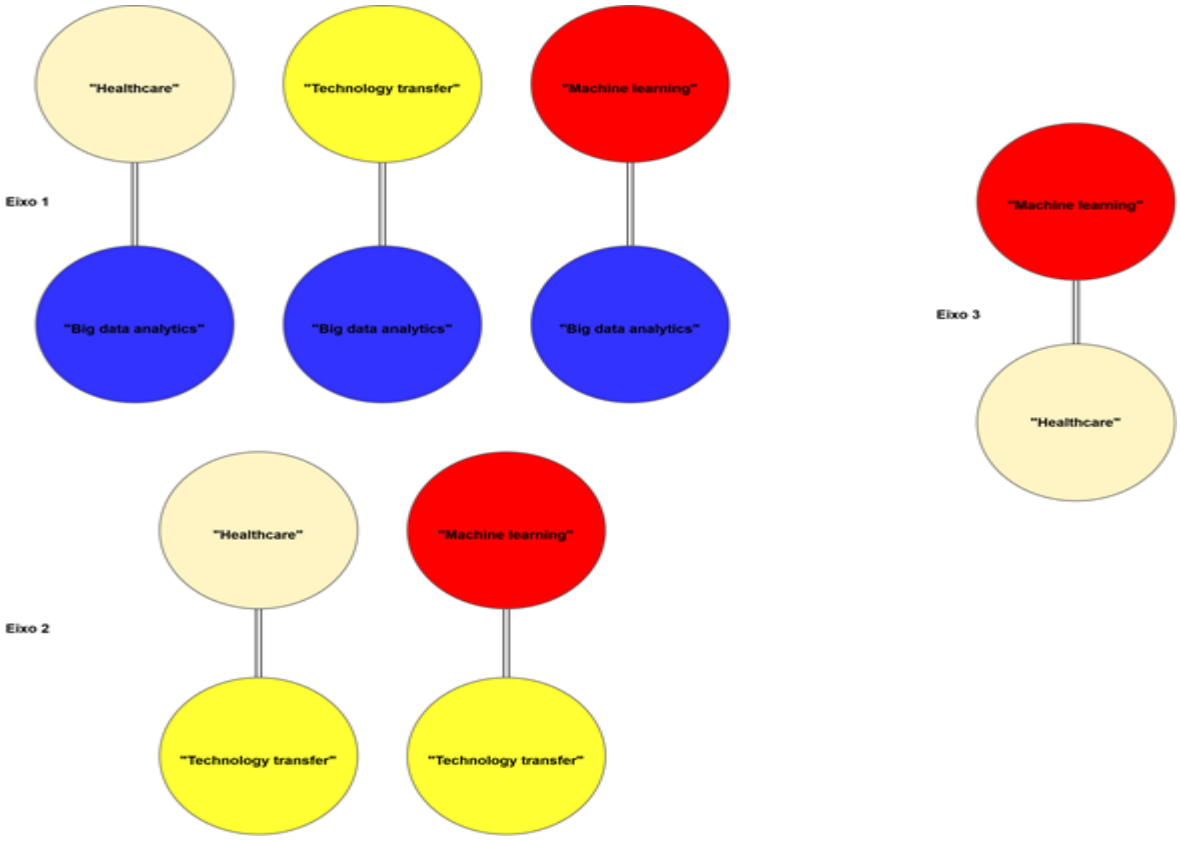
3.2 Etapas da pesquisa

3.2.1 Primeira etapa: revisão sistemática de literatura

A revisão sistemática de literatura foi desenvolvida, a fim de construir o referencial teórico necessário para entender o fenômeno inserido no problema de pesquisa. Vinculado a isto, a análise bibliométrica foi realizada através de artigos científicos, teses e livros onde a exploração, refinamento e revisitação foram executadas para melhor observação dos tópicos alinhados ao problema de pesquisa.

Devido à necessidade de estabelecer combinações das palavras-chave para melhor explorar os objetivos da pesquisa, fez-se necessário constituir três eixos de pesquisa, sendo que o eixo um visa compreender como ocorre a utilização da análise *big data* em cuidados de saúde; o eixo dois engloba a transferência de tecnologia ao contexto de cuidados em saúde, baseado em aprendizado de máquina por meio da análise *big data*; por fim, o eixo três explora a aplicação de aprendizado de máquina nos cuidados em saúde, conforme demonstrada na figura 12.

Figura 12 - Definição dos eixos de pesquisa



Fonte: Autoria própria (2022)

A metodologia definida para realização da revisão sistemática de literatura foi o *Methodi Ordinatio*, um método multicritério para tomada de decisão sobre a formação do portfólio de artigos de periódicos, visando compor o referencial teórico (PAGANI; KOVALESKI; RESENDE, 2015).

Conforme disposição das etapas do *Methodi Ordinatio* (fig. 13), inicialmente instituiu-se a intenção de pesquisa, composta pelas palavras-chave: *big data analytics, healthcare, technology transfer* e *machine learning,* organizadas em eixos de pesquisa conforme figura 12, estabelecendo a pesquisa definitiva necessária. Após o estabelecimento da intenção de pesquisa, decidiu-se utilizar as bases de periódicos *Web of Science*, *Scopus* e *Science Direct*, pela cobertura que respectivas bases promovem em relação a temas ligados à saúde, à computação e à tecnologia. Estabeleceu-se o período de 01/01/2000 a 31/12/2020, em decorrência da pesquisa exploratória desenvolvida e evolução das publicações, considerando título, *abstract* e palavras-chave.

Na pesquisa definitiva, 800 documentos foram encontrados de forma sistemática. Posteriormente, foram aplicados procedimentos de filtragem e

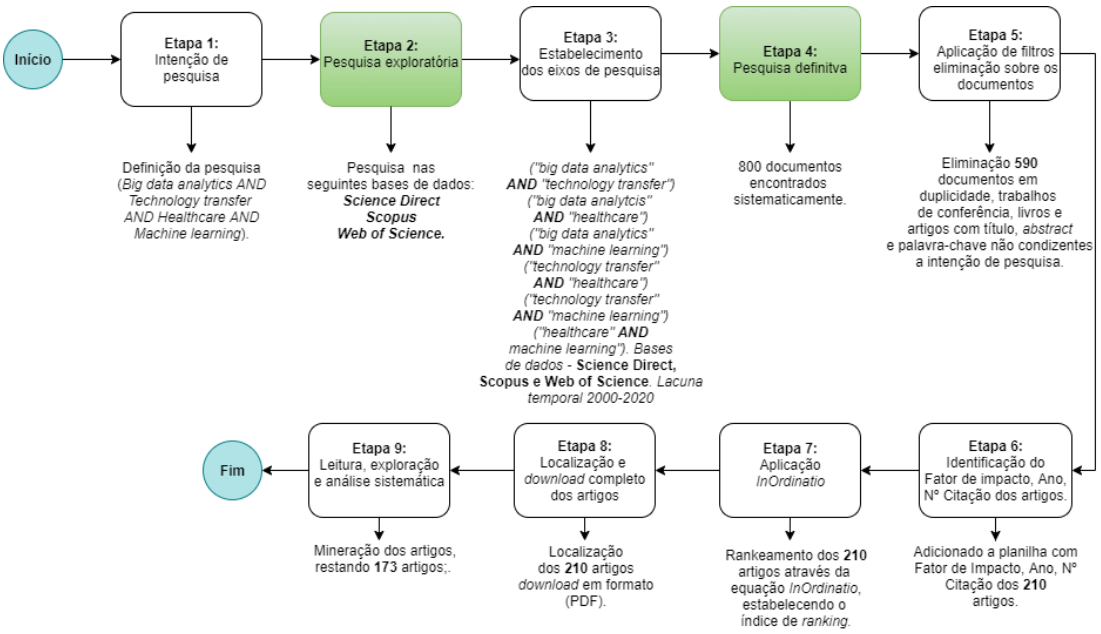
eliminação, sendo caracterizados pela duplicidade, trabalhos de conferência, livros e artigos com título, *abstract* e palavras-chave não condizentes com a intenção de pesquisa, apresentando resultado refinado em 590 documentos. O gerenciador de referências, Zotero, foi utilizado, a fim de organizar o portfólio constituído.

A próxima etapa foi realizada através da identificação do fator de impacto (FI)[1], número de citações e ano de publicação. Estabeleceu-se, por meio de uma planilha eletrônica, a aplicação da fórmula do *Methodi Ordinatio* (Apêndice A), conforme exposta abaixo:

$$InOrdinatio = (Fi/1000) + \alpha * [10 - (AnoPesq - AnoPub)] + (\Sigma Ci) \quad (13)$$

Após a aplicação e tratamento dos dados através da planilha eletrônica, 210 artigos foram localizados e realizados os devidos *downloads* em formato PDF; concluindo, assim, a fase de análise sistemática de literatura.

Figura 13 - Etapas de aplicação da metodologia *Methodi Ordinatio*



Fonte: Adaptado de Pagani, Kovaleski e Resende (2015)

Após a mineração dos artigos e o cumprimento de metas de leitura, foram utilizados 116 artigos nesta pesquisa, mais 6 artigos complementares, totalizando 122 artigos.

[1] JCR da *Clarivate Analytics*.

3.2.2 Segunda etapa: desenvolvimento do modelo preditivo com funcionalidades de análise *big data* e TT

Para compreender a solução proposta para o problema constituído nesta pesquisa, foi desenvolvido um modelo preditivo baseado na arquitetura de análise *big data* e TT em virtude da proporção de dados e informações em um formato contínuo e intenso – acerca de pacientes submetidos aos cuidados de saúde –, composto por cinco estágios inter-relacionados (fig. 14). Estágio 1, denominado captura de dados, considera diversas naturezas de dados. Por exemplo: dados clínicos, exames, prontuários eletrônicos médicos, prescrições diversas, dados de identificação, dentre outros.

Estágio 2, denominado processamento, este ocorre a partir de funções as quais preparam os dados para a etapa analítica. A função limpeza procura substituir os valores ausentes por valores médios. A função seleção identifica recursos que se correlacionam entre si, removendo atributos irrelevantes, aplicando a técnica baseada em correlação para determinar os atributos eficazes para o campo analítico e descoberta de conhecimento. Por fim, a função transformação aplica a escala variada de 0,0 a 1,0 utilizando a técnica mínimo-máximo para ocorrer a normalização de dados. Estágio 3 é onde são definidas as dimensões da população-alvo para aplicação da técnica preditiva, sendo:

- público-alvo – definição dos pacientes ou grupos de pacientes acometidos por doença;
- banco de dados – definição dos conjuntos de dados necessários para a análise preditiva;
- rótulo – resultado esperado sobre o público-alvo;
- conjunto de treinamento e teste – definição dos conjuntos de treinamento e teste para aplicação do algoritmo preditivo e medidas de desempenho.

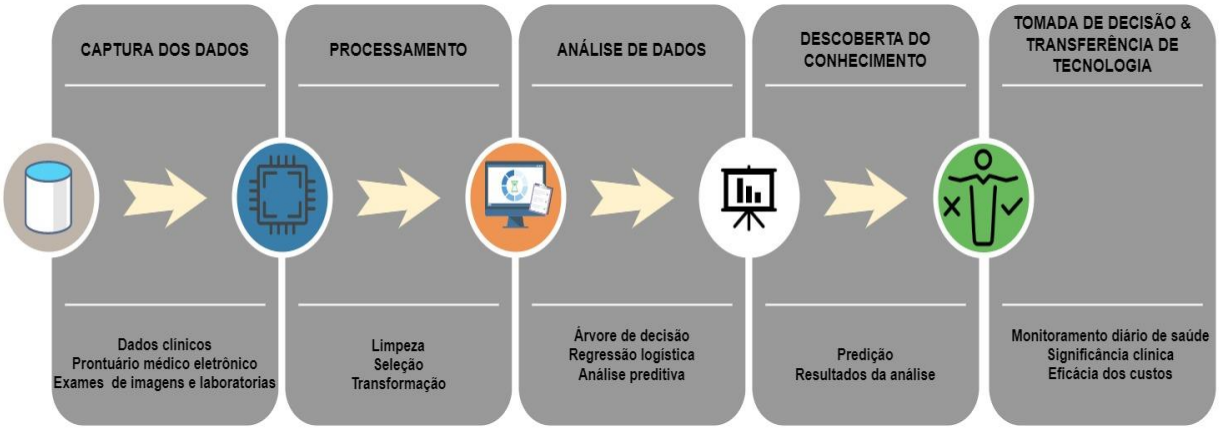
Nesse estágio, consideraram-se os algoritmos de aprendizado de máquina supervisionado, regressão logística binária e árvore de decisão, entendendo-se que a sua capacidade preditiva se torna adequada às investigações desta pesquisa (JOHNSTON *et al.,* 2019; HAQ *et al.,* 2020). Desta forma, ocorreu a aplicação do referido algoritmo a um conjunto de dados, o qual dividiu-se em 75% amostra para

treinamento e 25% da amostra para teste, possibilitando a avaliação de desempenho, considerando as medidas: acurácia, especificidade, sensibilidade, *recall*, precisão, curva ROC e AUC.

Estágio 4, denominado descoberta do conhecimento, com os resultados das predições realizadas no estágio 3, bem como as avaliações de desempenho produzidas sobre o algoritmo nos conjuntos de treinamento e teste, esses resultados são interpretados, cabendo à equipe médica a decisão do reconhecimento da significância clínica e modificação do processo de tratamento.

Estágio 5, descrito como tomada de decisão e transferência de tecnologia, tende a ser finalístico, onde a utilização dos resultados da pesquisa desenvolvidos no estágio 4 são utilizados, subsidiando a tomada de decisão e transferência de tecnologia com significância clínica, conforme demonstrado na figura 14.

Figura 14 - Modelo preditivo com funcionalidades de análise *big data* e transferência de tecnologia



Fonte: Autoria própria (2022)

3.2.2.1 Estágio 1 - Estabelecimento dos procedimentos de captura dos dados em repositório hospitalar

A definição do público-alvo é desenvolvida a partir dos critérios estabelecidos para repositório de pacientes submetidos em organizações de cuidados de saúde. Esta pesquisa considera uma doença como base na durabilidade das medidas de tratamento, acompanhamento e produção de dados ao longo do tempo. Na figura 15, são apresentados os critérios para exploração de dados no repositório.

Para o estabelecimento da população a ser utilizada no cálculo amostral, consideraram-se todos os municípios inseridos na região dos Campos Gerais no

estado do Paraná (quad. 4), compreendendo 19 municípios que realizam encaminhamentos para diagnóstico e tratamento na unidade hospitalar da Santa Casa de Misericórdia de Ponta Grossa (PR) que é a referência em tratamentos oncológicos na região.

Quadro 4 - Municípios pertencentes à região dos Campos Gerais do Paraná e sua população estimada

Município considerado pertencente à região dos Campos Gerais	População estimada (IBGE 2020).
Arapoti	28.300
Castro	71.809
Carambeí	23.825
Curiúva	15.196
Imbaú	13.282
Ipiranga	15.087
Ivaí	13.965
Jaguariaíva	35.027
Ortigueira	21.960
Palmeira	33.994
Piraí do Sul	25.617
Ponta Grossa	355.336
Porto Amazonas	4.874
Reserva	26.825
São João do Triunfo	15.241
Sengés	19.385
Telêmaco Borba	79.792
Tibagi	20.607
Ventania	12.088
Total de Municípios = 19	**População total estimada = 832.210**

Fonte: Autoria própria (2022)

Para o cálculo amostral, foi apurado por meio da análise documental que 3.095 pacientes foram atendidos pelo serviço de oncologia do hospital Santa Casa de Misericórdia - Ponta Grossa (PR). Dessa forma, a totalidade da população de pacientes resulta em um valor mínimo aceitável para exploração da amostra em repositório hospitalar. Dados de cada paciente foram orientados pela estrutura de preditores preestabelecida, onde foram submetidos ao modelo preditivo desenvolvido, não havendo intervenção, tendo em vista que a validação ocorreu por intermédio de

medidas de desempenho sobre os algoritmos aplicados. Na figura a seguir, são apresentados os critérios para exploração de dados.

Figura 15 - Critérios para exploração de dados no repositório

Fonte: Autoria própria (2022)

Os dados inseridos nos prontuários médicos e nos Registros Hospitalares de Câncer (RHC) representam centros de coleta de forma sistemática e contínua acerca de informações sobre pacientes atendidos em uma unidade hospitalar com diagnóstico inicial de câncer. As principais funções de tal RHC são clínicas, considerado um recurso para acompanhar e avaliar a qualidade do trabalho desenvolvido nos hospitais, incluindo resultados sobre o tratamento do câncer. Desta forma, estes dados são coletados nos prontuários médicos e RHC de pacientes que atendem aos critérios de inclusão. Uma estrutura de preditores é desenvolvida, objetivando a sistematização necessária para obtenção dos resultados esperados: sendo tal sistematização apresentada a seguir:

Quadro 5 - Descrição dos preditores do banco de dados

Variável	Descrição
y	Retornou ou não ao tratamento
x1	Diagnóstico e tratamento anteriores
x2	Ano
x3	Recno - identificador único e sequencial
x4	Tipo de caso
x5	Sexo
x6	Idade
x7	Local do nascimento
x8	Raça / cor
x9	Escolaridade
x10	Clínica do primeiro atendimento
x11	Clínica de início do tratamento

x12	Histórico familiar de câncer
x13	Histórico de consumo de bebida alcoólica
x14	Histórico de consumo de tabaco
x15	Estado de residência
x16	Município de residência
x17	Ano de diagnóstico
x18	Origem do encaminhamento
x19	Exames relevantes para o diagnóstico e planejamento da terapêutica do tumor
x20	Estado civil
x21	Ano da triagem
x22	Ano da primeira consulta
x23	Base mais importante para o diagnóstico do tumor
x24	Localização primária (Categoria 3d)
x25	Localização primária detalhada (Subcategoria 4d)
x26	Tipo histológico do tumor primário
x27	Lateralidade do tumor
x28	Codificação do estádio clínico segundo classificação TNM
x29	Estadiamento clínico do tumor (TNM)
x30	Outros estadiamentos clínicos do tumor
x31	Ptnm - Classificação histopatológica pós-cirúrgica
x32	Principal razão para a não realização do tratamento antineoplásico no hospital
x33	Ano do início do primeiro tratamento específico para o tumor no hospital
x34	Primeiro tratamento recebido no hospital
x35	Estado da doença ao final do primeiro tratamento no hospital
x36	Número do CNES do hospital
x37	UF da unidade hospitalar
x38	Município da unidade hospitalar
x39	Ocupação principal
x40	Data triagem
x41	Data que começou o tratamento

Fonte: Autoria própria (2022)

Em posse dos dados, conforme estrutura de preditores, foram definidos dois conjuntos de dados: um para treinamento (75%), e outro para teste (25%), onde encontrou-se a taxa de ressubmissão oncológica.

3.2.2.2 Estágio 2 - Definição das funcionalidades de processamento e normalização dos dados

O estágio 2 se encontra situado entre o estágio 3 e o estágio 1: O estágio 1 compreende na captação dos dados em que os conjuntos de dados armazenados

consistem em valores ausentes, ruídos e registros inconsistentes que necessitam ser processados com eficácia antes de uma análise aprofundada; o estágio 3 por sua vez consiste na aplicação do aprendizado de máquina supervisionado (ABDELAZIZ *et al.,* 2018; JOHNSTON *et al.,* 2019; CHAUHAN; KAUR; CHANG, 2020).

Para o processamento dos dados, são considerados fases: limpeza dos dados, transformação dos dados e seleção dos dados. Um detalhamento será apresentado a seguir.

Limpeza dos dados: objetiva remover os valores ausentes e substituí-los por valores médios para obter a recuperação de padrões eficazes e eficientes; inicialmente valores ausentes são substituídos por valores *null*, ou seja, valor desconhecido, inserindo manualmente os valores. Outra técnica adequada está em remover os valores com média e mediana em relação ao conjunto, observando as discrepâncias dos valores (CHAUHAN; KAUR, CHANG, 2020).

Transformação dos dados: considerada uma técnica com faixas de escala variada de 0,0 a 1,0, objetivando transformar os dados. Possui várias possibilidades de classificação e agrupamento. Tal transformação é realizada através da técnica mín-máx, em que redimensiona linearmente cada recurso para intervalo de 0,0 a 1,0 (ABDELAZIZ *et al.,* 2018). Esta técnica e calculada a partir da seguinte fórmula:

$$Z = \frac{x - min(x)}{[max(x) - min(x)]} \quad (14)$$

Seleção de dados: consiste em remover os atributos irrelevantes e selecionar os recursos mais apropriados que correlacionam entre si. É considerada uma etapa obrigatória (ANMED *et al.,* 2020; CHAUHAN; KAUR; CHANG, 2020). Para selecionar adequadamente, é estabelecida a técnica baseada em correlação entre os preditores, a qual é definida através da forma linear ou não linear do conjunto de dados; entendendo-se que os recursos mais correlacionados com os valores das classes em comparação uns com os outros são os melhores resultados. A principal intenção é reduzir a dimensionalidade dos dados para a etapa seguinte (AHMED *et al.*, 2020; CHAUHAN; KAUR; CHANG, 2020). Após a realização do processamento, naturalmente, estes dados estão preparados para uma investigação aprofundada através da aplicação de aprendizado de máquina.

3.2.2.3 Estágio 3 - Aplicação do algoritmo aos conjuntos de dados e avaliação de desempenho

Neste estágio, são aplicados os algoritmos de regressão logística e árvore de decisão ao conjunto de dados, e a definição do resultado esperado ocorre de forma preditiva, conhecida como rótulo. Representa a resposta esperada, ou seja, a probabilidade de o paciente retornar ou não aos serviços de oncologia. A atividade seguinte considerou o estabelecimento dos preditores a partir dos registros produzidos durante o período de apuração, sendo os candidatos baseados nos critérios de definição da população-alvo.

A área sob a curva (AUC), é considerada um procedimento de reamostragem capaz de selecionar o valor ótimo do hiperparâmetro. A AUC minimiza o espaço otimista, por isso é considerada uma medida de avaliação de desempenho que demonstra a acurácia do desempenho do modelo de regressão logística binária (JOHNSTON *et al.,* 2019). Durante a aplicação da validação cruzada, constituiu-se a avaliação do modelo no conjunto de validação. Este processo ocorre de forma interativa em cada conjunto, objetivando fornecer um desempenho de validação cruzada estimado da regressão logística com variância específica.

3.2.2.4 Estágio 4 - Descoberta do conhecimento por meio da apuração dos resultados da análise

Para este estágio, são apurados os resultados da análise aplicada no estágio anterior de forma que a descoberta de conhecimento ocorre a partir da predição dos resultados encontrados, subsidiando o estágio subsequente (ZHANG *et al.,* 2019). Os resultados são apresentados a partir das medições de desempenho aplicadas no conjunto de treinamento e teste, podendo ser interpretadas e sintetizadas para a tomada de decisão e TT em formatos de hipóteses as quais contribuem para otimização e adaptação clínica.

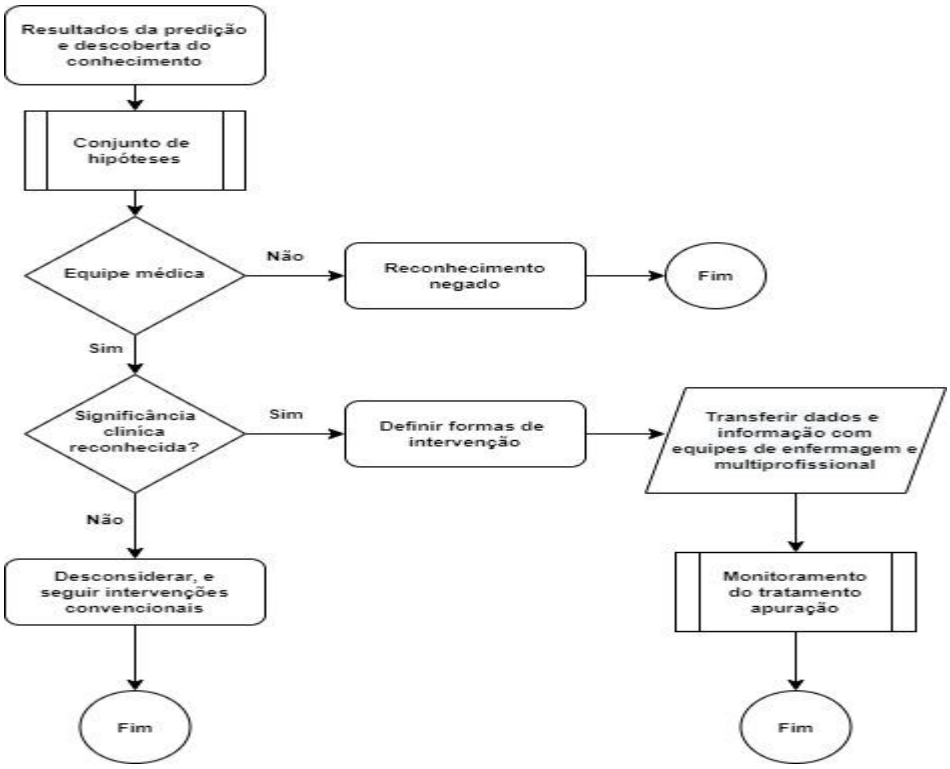
3.2.2.5 Estágio 5 - Subsídio à tomada de decisão e transferência de tecnologia

Os resultados gerados e interpretados são utilizados para adaptação clínica e otimização dos procedimentos adotados. No entanto é necessário considerar os

pontos de vista das partes interessadas, a saber: pacientes objetivam o restabelecimento da sua condição de recuperação; o médico almeja o conjunto de hipóteses geradas pela ferramenta, buscando a eficácia da personalização do acompanhamento e tratamento quando a população-alvo se beneficia com o atingimento de metas; e o sistema de saúde, enquanto pagador, justificando os subsídios financeiros aplicados ao tratamento.

A capacidade de interpretação dos resultados exige treinamento acerca dos relatórios produzidos pela ferramenta. Os médicos devem ser capazes de compreender o que as predições representam. A aceitação da predição depende da decisão médica em considerar os resultados. A tomada de decisão e TT em nível clínico (fig. 16) permite aos médicos: orientarem a atuação de equipes de profissionais de enfermagem e multiprofissional, com base nas predições realizadas; aproximarem a ação supervisora do responsável, desenvolverem avaliação de desempenho conjuntamente com o monitoramento diário dos cuidados em saúde, conectarem-se os aspectos comuns inerentes às partes interessadas.

Figura 16 - Fluxograma do processo de tomada de decisão clínica e TT



Fonte: Autoria própria (2022)

3.2.3 Terceira Etapa: aplicação das medidas de desempenho para validação do algoritmo regressão logística e árvore de decisão

Diferentes métricas são aplicadas para a medição de desempenho dos algoritmos: regressão logística e árvore de decisão. Para calcular as medidas de desempenho – índice Kappa, Acurácia, especificidade, sensibilidade, curva ROC, AUC –, foi desenvolvida a matriz de confusão binária (JOHNSTON *et al.,* 2019; HAQ *et al.,* 2020; LI *et al.,* 2020).

A matriz de confusão binária estabelece a saída prevista como verdadeiro positivo (*true positive TP*) quando o indivíduo está acometido por câncer e apresenta piora; para verdadeiro negativo (*true negative TN*) quando o indivíduo saudável é classificado como saudável. Falso positivo (*false positive FP*) se um indivíduo saudável é considerado um sujeito com piora; para falso negativo (*false negative FN*), se o indivíduo com câncer for considerado um indivíduo saudável (HAQ *et al.,* 2020).

Com base nas matrizes de confusão binária, as medidas de desempenho são calculadas conforme a seguir: índice Kappa (k) - capaz de expressar a confiabilidade do teste através do índice k indicador de concordância ajustada, pois leva em consideração a concordância, devida à chance. A tabela 1 destaca os níveis de concordância.

Tabela 1 - Níveis de concordância índice Kappa

Kappa	Concordância
<0,00	Nenhuma
0,00-0,20	Fraca
0,21-0,40	Sofrível
0,41-0,60	Regular
0,61-0,80	Boa
0,81-0,99	Ótima
1,00	Perfeita

Adaptado de Landis & Koch (1977)

O índice Kappa é estimado como:

$$k = \frac{P_O - P_e}{1 - P_e} \quad (14)$$

Sendo:

P_O: Proporção de concordâncias observadas;

P_e: Proporção de concordâncias esperadas.

Acurácia (Acc): descreve o desempenho geral do classificador; sua notação matemática é apresentada a seguir:

$$AC = \frac{a+d}{a+b+c+d} \quad (15)$$

Sensibilidade: demonstra que o teste-diagnóstico é positivo e que a pessoa tem a doença; também chamado de Taxa de Positivo Verdadeiro (TPR) e calculado a partir da seguinte equação-matemática:

$$s = P(T = +|D = 1) = \frac{a}{n_1} = \frac{a}{a+c} \quad (16)$$

Especificidade: apresenta como um teste preditivo; é negativo, e a pessoa é saudável. Desta forma, a especificidade é expressada pela equação abaixo:

$$e = P(T = -|D = 0) = \frac{d}{n_2} = \frac{d}{b+d} \quad (17)$$

Outra medida de desempenho utilizada é a curva de características operacionais do receptor (ROC), baseada em um conjunto de ferramentas gráficas que apresenta uma análise comparativa entre a taxa de verdadeiro positivo e a taxa de falso positivo sobre o desempenho do algoritmo, e a curva de área sob a curva (AUC) caracteriza a curva ROC através da medida da área, em que um alto valor representa um desempenho aceitável ao modelo (JOHNSTON *et al.,* 2019; HAQ *et al.,* 2020).

3.3 Procedimentos de coleta e análise de dados

Nesta pesquisa, consideraram-se os dados que se encontram no banco de dados da instituição hospitalar: dados do prontuário eletrônico, dados de exames laboratoriais, de imagens e dados inseridos no RHC. A forma de apresentação é através de relatórios e planilhas.

Os dados são de um conjunto de pacientes acometidos por câncer: os quais necessitam de acompanhamento médico para o tratamento. Tais dados foram levantados com base nos critérios de seleção e preditores preestabelecidos. Para realização do estágio 2 da ferramenta, foi utilizado o pacote *Dplyr* no *software Rstudio*@ para efetuar a limpeza, transformação e seleção dos dados.

Para desenvolver o estágio 3, “aplicação dos algoritmos”, e o estágio 4, “descoberta do conhecimento”, foram utilizados os pacotes, *caret, rpart.plot* (REPS *et al.,* 2018), com código aberto em linguagem R e *Python*; a aplicação dos pacotes foi utilizada nos *softwares R*@, versão 4.2.0 e *Rstudio*@ que permitiram desenvolver conjuntamente as medidas de desempenho sobre o modelo preditivo.

4 ANÁLISE E DISCUSSÃO DOS DADOS

Neste capítulo, encontram-se todos os procedimentos, resultados e informações dos pacientes submetidos a tratamento oncológico entre 2010 e 2019. Esses dados foram obtidos por meio de acesso ao banco de dados e sistema eletrônico de gestão em saúde *Tasy* utilizado pelo hospital Santa Casa de Misericórdia de Ponta Grossa (PR). O projeto se encontra aprovado no referido hospital e no Comitê de Ética e Pesquisa da Universidade Tecnológica Federal do Paraná sob o CAAE: 50596221.7.0000.5547 e parecer 4.993564. Todas as informações estão organizadas conforme as etapas prescritas no modelo preditivo proposto, em termos de passos metodológicos.

4.1 Elaboração do *dataset* para construção do modelo preditivo

Totalizando 3.095 pacientes atendidos entre os anos de 2010 e 2019, pelo serviço de oncologia do hospital Santa Casa de Misericórdia, localizado no município de Ponta Grossa (PR). Hospital este fundado em 1912, marcado por uma gestão filantrópica e aderência significativa ao sistema único de saúde pública. A Santa Casa de Ponta Grossa representa uma referência na região dos campos gerais em tratamento de oncologia. Classificada pelo Ministério da Saúde como Unidade de Alta Complexidade em Oncologia, este complexo hospitalar possui capacidade assistencial e tecnológica para diagnóstico e tratamento de pacientes com câncer.

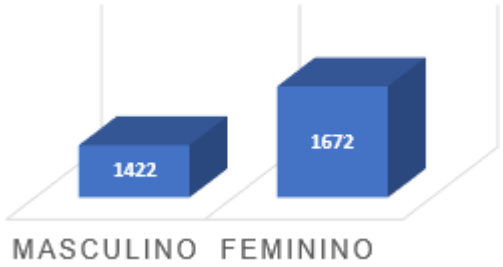
O banco de dados deste trabalho consiste em uma série histórica de 2010 a 2019, com 41 variáveis, que está disponível no Sistema de Registro Hospitalar de Câncer (SisRHC) e no Sistema de Gestão de Saúde *Tasy*. Possui 3.095 pacientes: sendo 582 (18,80%) do ano de 2010; 520 (16,80%) do ano de 2011; 403 (13,02%) do ano de 2012; 420 (13,57%) do ano de 2013; 391 (12,63%) do ano de 2014; 66 (2,13%) do ano de 2015; 163 (5,27%) do ano de 2016; 207(6,69%) do ano de 2017; 6 (0,19%) do ano de 2018; 337 (10,89%) do ano de 2019. Foram criadas duas amostras aleatórias, sendo uma com 2.301 (75%) e uma com 744 (25%) dos pacientes.

A composição das variáveis foi definida na metodologia e ampliada conforme possibilidades de acesso, ou seja, aquelas que estavam inseridas no sistema

eletrônico de gestão de saúde *Tasy* e no banco de dados disponibilizado pela equipe de tecnologia da informação do referido hospital.

A análise descritiva do banco de dados foi desenvolvida com o intuito de compreender alguns dados por meio de visualização gráfica, considerando a caracterização dos pacientes, procedimentos adotados em ambiente hospitalar e informações complementares sobre estilo de vida (gráf. 1). Ele representa a população em termos de gênero:

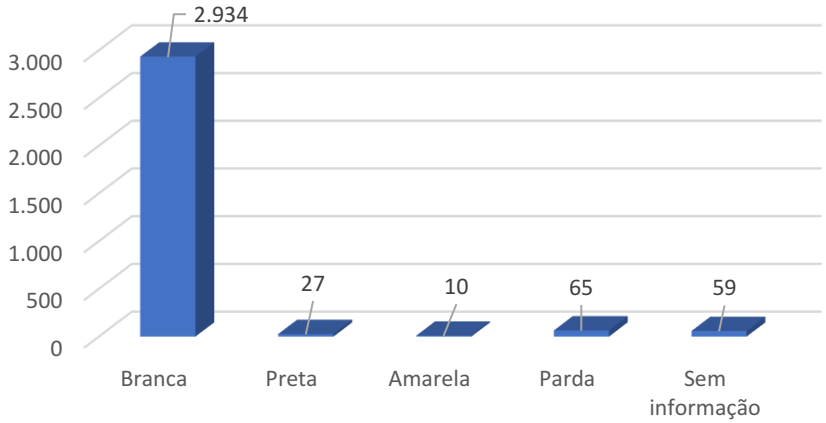
Gráfico 1 - Gênero dos pacientes



Fonte: Autoria própria (2022)

No próximo gráfico, segundo os pacientes, 2.934 (94,80%) consideraram a cor da sua pele, branca; 27 (0,87%), preta; 10 (0,32%), amarela; 65 (2,10%), parda; 59 (1,91%) não informaram nenhuma cor.

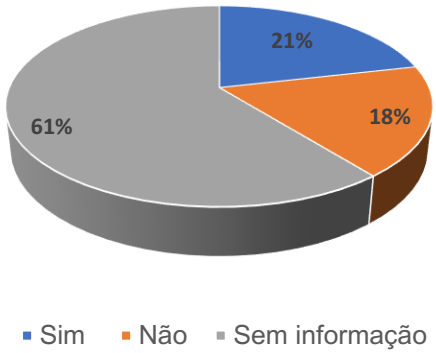
Gráfico 2 - Raça/cor dos pacientes



Fonte: Autoria própria (2022)

Na variável histórico familiar, os pacientes que têm histórico de câncer são 663 (21,42%), os que não têm são 551 (17,80%) e os pacientes sobre os quais não há informação, 1881 (60,78%). Os dados são representados pelo gráfico a seguir.

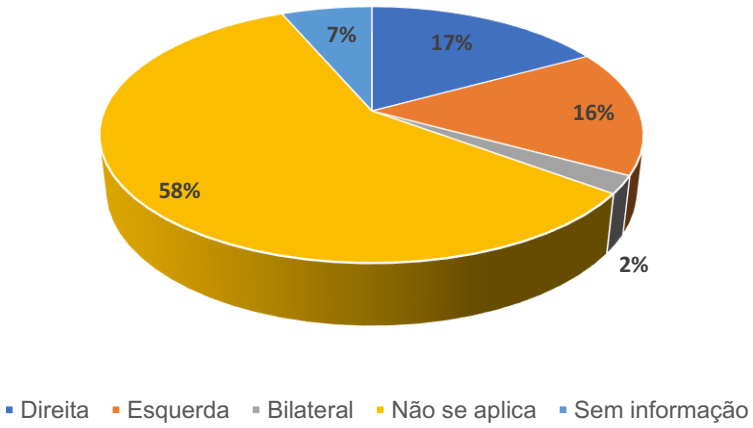
Gráfico 3 - Histórico familiar de câncer para os pacientes



Fonte: Autoria própria (2022)

Lateralidade do tumor apresentada no gráfico seguinte, dados obtidos por meio dos exames de diagnóstico do tumor primário, tem 527 pacientes que correspondem (17,03%) com tumores no lado direito, 492 (15,90%) com tumores no lado esquerdo, 69 (2,23%) com tumores bilaterais, 1.808 (58,42%) não se aplica e 199 (6,43%) sem informação.

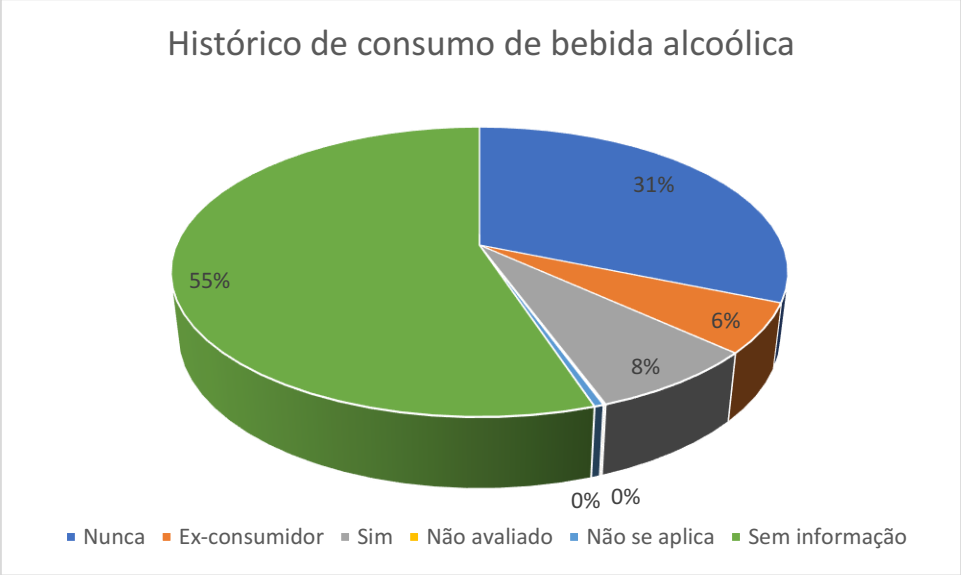
Gráfico 4 - Lateralidade do tumor dos pacientes



Fonte: Autoria própria (2022)

Quanto ao histórico de consumo de bebida alcoólica dos pacientes, 1.704 (55,06%) não informaram, 968 (31,28%) nunca consumiram, 236 (7,63%), sim, 171 (5,53%) ex-consumidores, 13 (0,42%) não se aplica, 3 (0,10%) não avaliados. Dados são apresentados no gráfico 5.

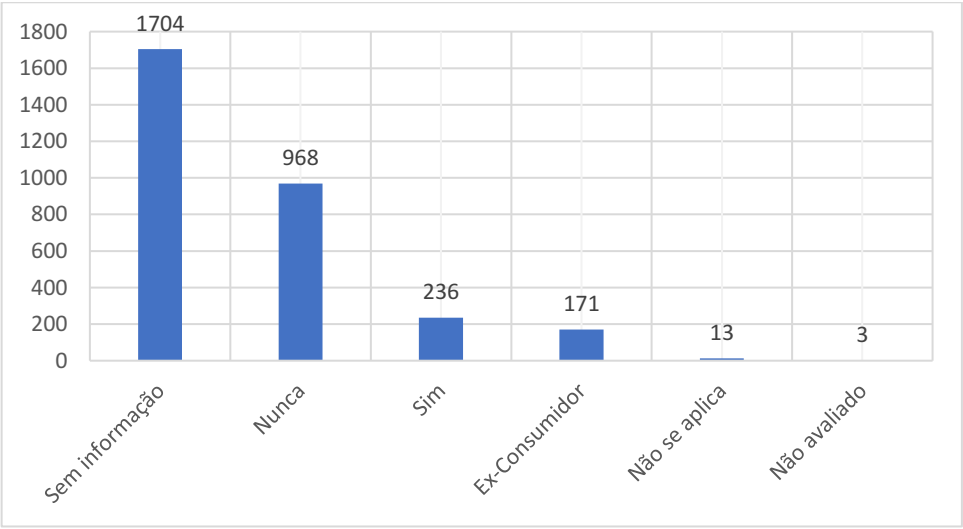
Gráfico 5 - Histórico de consumo de bebida alcoólica dos pacientes



Fonte: Autoria própria (2022)

Para o consumo de tabaco, destacado no gráfico 6, 1.141 (36,87%) pacientes não informaram o uso ou não de tabaco, 1.000 (32,31%) nunca usaram, 593 (19,16%), sim, 351 (11,34%) ex-consumidores, 9 (0,29%) não se aplica, 1 (0,03%) não avaliado.

Gráfico 6 - Histórico de consumo de tabaco dos pacientes

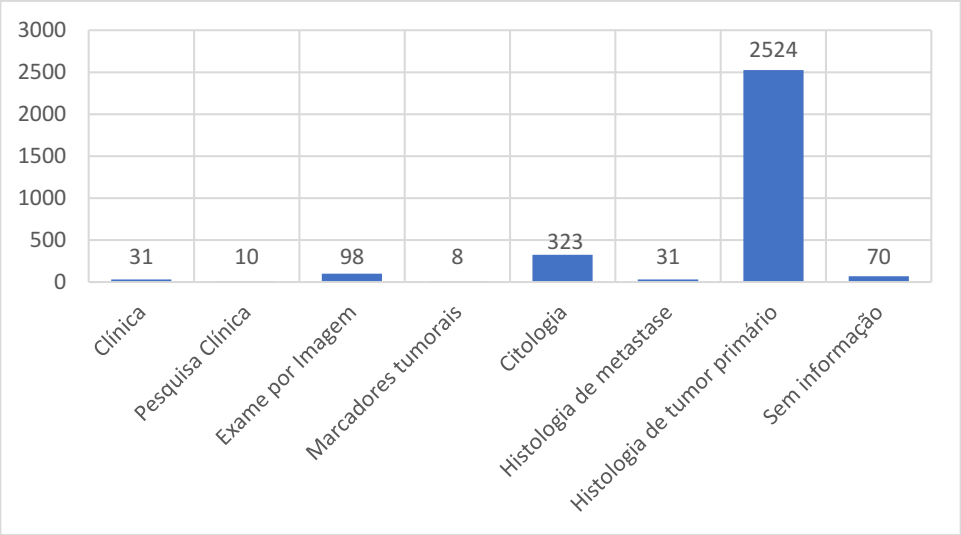


Fonte: Autoria própria (2022)

Base importante para diagnóstico do tumor. Consideram-se todas as possibilidades de exames realizados. Da população dos pacientes representada, 31 (1,00%) realizam exames clínicos; 10 (0,32%), pesquisa clínica; 98 (3,16%), exames por imagem; 8 (0,25%), marcadores tumorais; 323 (10,43%), citologia; 31 (1,00%),

histologia de metástase; 2.524 (81,55%), histologia de tumor primário; 70 (2,26%), não informaram qual meio de diagnóstico foi utilizado; identificado como sem informação. O gráfico 7 destaca esses quantitativos de maneira visual e comparativa.

Gráfico 7 - Base importante para diagnóstico do tumor



Fonte: Autoria própria (2022)

Compreender as classificações utilizadas para cada tipo de câncer, aplicada a cada paciente, foi necessário para caracterizar o primeiro diagnóstico e o primeiro tratamento. O quadro 6 apresenta códigos da Classificação Internacional de Doenças - CID-10 e sua descrição. O CID-10 foi utilizado devido ao período de apuração dos pacientes, destacando sua vigência.

Quadro 6 - Códigos CID-10 e sua descrição

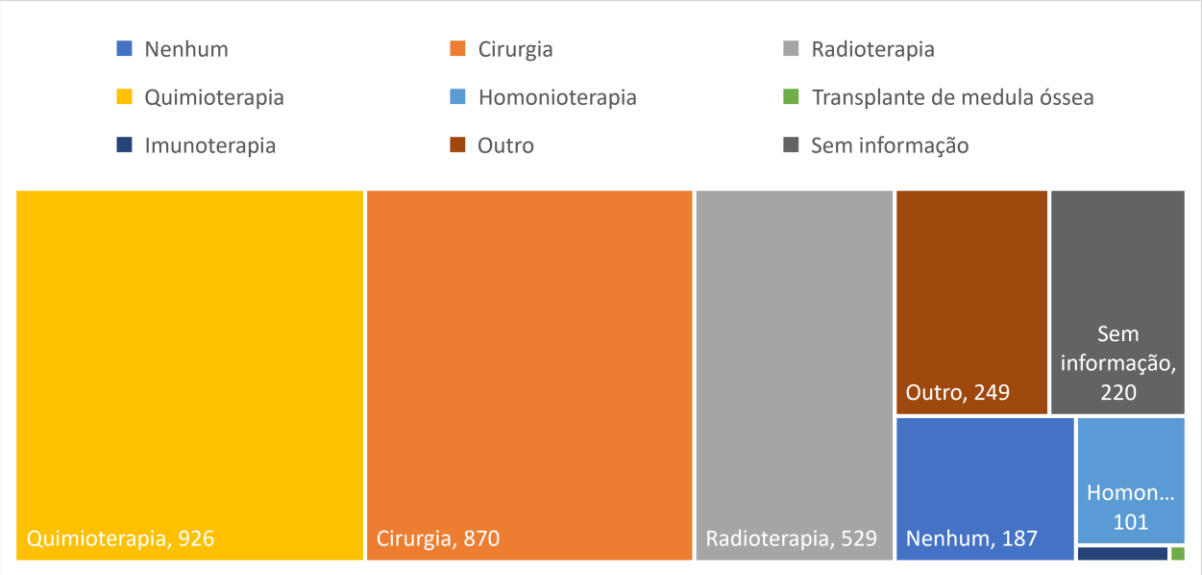
Código	Descrição do código
c00	Lábio
c01	Base da língua
c02	Língua
c03	Gengiva
c04	Assoalho da boca
c05	Palato
c06	Boca: lugares não especificados
c07	Glândula parótida
c09	Amígdala
c10	Orofaringe
c11	Nasofaringe
c13	Hipofaringe
c14	Lábio, cavidade oral e faringe
c15	Esôfago
c16	Estômago

c17	Intestino delgado
c18	Cólon
c19	Junção retossigmoide
c20	Reto
c21	Ânus e canal anal
c22	Fígado e ductos biliares intra-hepáticos
c23	Vesícula biliar
c24	Outras partes não especificadas no trato biliar
c25	Pâncreas
c31	Seios e acessórios
c32	Laringe
c34	Pulmão
c38	Coração, mediastino e pleura
c39	Sistema respiratório e órgãos intratorácicos
c40	Ossos e cartilagem
c41	Ossos e cartilagem
c43	Melanoma maligno de pele
c44	Pele
c47	Nervos periféricos e sistema nervoso autônomo
c48	Retroperitônio e peritônio
c49	Tecidos conjuntivos moles
c50	Mama
c51	Vulva
c52	Vagina
c53	Colo do útero
c54	Corpo uterino
c55	Útero e partes não especificadas
c56	Ovário
c58	Placenta
c60	Pênis
c61	Próstata
c62	Testículo
c64	Rim, exceto pelve renal
c67	Bexiga
c69	Olhos e anexos
c71	Cérebro
c72	Medula espinhal
c73	Glândula tiroide
c74	Glândula adrenal
c76	Maus definidos
c77	Secundária não especificada: linfonodos
c80	Sem especificação

Fonte: Autoria própria (2022)

Próxima representação gráfica considera as modalidades de tratamento aplicadas à população de pacientes submetidos a atendimento no serviço de oncologia. Proporções consideram uma parcela significativa de casos que são tratados pela modalidade terapêutica: quimioterapia, cirurgia oncológica e radioterapia.

Gráfico 8 - Quantitativo de modalidades terapêuticas aplicadas entre 2010 e 2019



Fonte: Autoria própria (2022)

Na próxima seção, a aplicação dos algoritmos de aprendizado de máquina supervisionado é desenvolvida no *dataset bd_oncology*.

4.2 Aplicação do modelo preditivo no *dataset bd_oncology*: possibilidades de algoritmos de aprendizagem supervisionada

4.2.1 Algoritmo regressão logística aplicado ao *dataset bd_oncology*

O *dataset bd_oncology* constituído no aplicativo Microsoft Excel 2019@, em formato separado por vírgulas, permitiu manusear a importação ao *software* de análise de dados *Rstudio*@. A configuração foi simplificada para o melhor manejo dos dados, os quais foram estruturados de maneira relacional e considerando os preditores pré-selecionados. Inicialmente é apresentado o *script* desenvolvido no *software Rstudio*@.

Script* 1 - Regressão logística aplicada ao *bd_oncology* no *software Rstudio

```
RM(LIST = LS())
```

```
LIBRARY(TCLTK) # TCL/TK PARA ABRIR O BD
BD<-READ.CSV2("E:/MYLLER/BANCOS/BD25.CSV",HEADER=T)
ATTACH(BD)
NAMES(BD)
#VARIAVEIS#
Y<-BD$RETORNOU.OU.NÃO
X1<-BD$DIAGANT
X2<-BD$ANO
X3<-BD$RECNO
X4<-BD$TPCASO
X5<-BD$SEXO
X6<-BD$IDADE
X7<-BD$LOCAL777S
X8<-BD$RACACOR
X9<-BD$INSTRUC
X10<-BD$CLIATEN
X11<-BD$CLITRAT
X12<-BD$HISTFAMC
X13<-BD$ALCOOLIS
X14<-BD$TABAGISM
X15<-BD$ESTADRES
X16<-BD$PROCEDEN
X17<-BD$ANO18IDI
X18<-BD$ORIENC
X19<-BD$EXDIAG
X20<-BD$ESTCONJ
X21<-BD$ANTRI
X22<-BD$DT18ICON
X23<-BD$BASMAIMP
X24<-BD$LOCTUDET
X25<-BD$LOCTU18I
X26<-BD$TIPOHIST
X27<-BD$LATERALI
X28<-BD$TNM
X29<-BD$ESTADIAM
```

```
X30<-BD$OUTROESTA
X31<-BD$PTNM
X32<-BD$RZNTR
X33<-BD$DTINITRT
X34<-BD$A18ITRATH
X35<-BD$ESTDFIMT
X36<-BD$CNES
X37<-BD$UFUH
X38<-BD$MUUH
X39<-BD$OCUPACAO
X40<-BD$D_TRI_DIAG
X41<-BD$D_1CONS_TRAT
#REGRESSÃO LOGISTICA
#75 POR CENTO
MODELO1<GLM(Y~X1+X2+X3+X4+X5+X6+X7+X8+X9+X10+X11+X12+X13+X14+X15+X16+X17+X
18+X19+X20+X21+X23+X24+X25+X26+X27+X28+X29+X30+X31+X32+X33+X34+X35+X39+X40+X
41,FAMILY=BINOMIAL(LINK="LOGIT"))
SUMMARY(MODELO1)
MODELOII<-GLM(Y~X1+X2+X4+X10+X11+X13+X14+X17+X23+X24+X25+X28+X31
+X35+X39+X40+X41,FAMILY=BINOMIAL(LINK="LOGIT"))
SUMMARY(MODELOII)
#25 POR CENTO
MODELO2<-GLM(Y~X1+X2+X3+X4+X5+X7+X8+X9+X10+X11+X12+X13+X14+X16+X17+X18
+X19+X20+X21+X22+X23+X24+X25+X26+X28+X29+X30+X31+X32+X33+X35+X36+X37+
X38+X39+X40+X41,FAMILY=BINOMIAL(LINK="LOGIT"))
#SUMMARY(MODELO2)
MODELOIII<-GLM(Y~X1+X4+X10+X11+X26+X31+X35+X39,
FAMILY=BINOMIAL(LINK="LOGIT"))
SUMMARY(MODELOIII)
```

Fonte: Autoria própria (2022)

Foram realizados dois modelos de regressão logística, um para cada amostra, ou seja, 75% para treinamento e 25% para teste, separados de maneira aleatória. Utilizou-se o procedimento de validação cruzada, na identificação das variáveis. A variável dependente considerada foi o preditor: retornou ou não para fazer o

tratamento. O *script* detalhado abaixo destaca como o modelo generalizado foi desenvolvido com todas as variáveis pré-selecionadas.

***Script* 2 - Modelo generalizado de regressão logística**

```
modelo<-glm(y~x1+x2+x3+x4+x5+x6+x7+x8+x9+x10+x11+x12+x13+x14+x15+x16+x17+x18
+x19+x20+x21+x23+x24+x25+x26+x27+x28+x29+x30+x31+x32+x33+x34+x35+x39+x40+x41,
family=binomial(link="logit"))
```

Fonte: Autoria própria (2022)

Para ajustar o modelo, foram calculados o desvio nulo e o desvio residual: o primeiro representa a análise de todos os parâmetros, no caso do modelo saturado; o segundo tipo de desvio destaca a análise apenas dos parâmetros do modelo, pondo-se como uma medida de ajuste. O desvio nulo (2320 graus de liberdade) próximo do desvio residual (2283), para amostra de 75% da população e amostra de 25% desvio nulo (773 graus de liberdade), próximo do desvio residual (736); com esse desvio indica que os modelos foram adequados. A tabela 2 apresenta os ajustes dos modelos conforme as amostras selecionadas, e a tabela 3 destaca a estatística descritiva do modelo.

Tabela 2 - Ajuste dos modelos segundo as amostras

Amostra	Tipo de desvio	Desvio	Graus de Liberdade	AIC
75%	Desvio nulo	2302,80	2320	1544,9
	Desvio residual	1468,90	2283	
25%	Desvio nulo	794,17	773	579,63
	Desvio residual	503,63	736	

Fonte: Autoria própria (2022)

O teste de *Hosmer-Lemeshow*, utilizado para medir o ajuste de um teste estatístico (ABE *et al.,* 2021), expondo que não houve diferença estatisticamente significativa entre as classificações para amostra de 75% (com todas as variáveis do modelo), apresentou os seguintes resultados: X^2=20,463, df = 8 e um p-valor = 0,008718, como o p-valor abaixo de 0,05. Os dados do modelo de regressão logística não estão ajustados segundo o modelo de regressão logística.

. O teste de *Hosmer-Lemeshow* (apenas com as variáveis significativas do modelo anterior): X^2=20,463, df = 8 e um p-valor = 0,008718, como o p-valor abaixo de 0,05. Os dados do modelo de regressão logística não estão ajustados conforme do modelo de regressão logística.

Tabela 3 - Estatística descritiva do modelo

	Mínimo	1Q	Mediana	3Q	Máximo

75%	-32.769	0.1413	0.2747	0.4551	26.123
25%	309.318	0.09242	0.26558	0.48339	205.752

Fonte: Autoria própria (2022)

Foram realizados os ajustes dos modelos conforme as amostras e apurando sua estatística descritiva. Foi desenvolvido um modelo de regressão logística, com a variável dependente e com as variáveis que foram significantes ao modelo anterior para cada amostra de 75% e 25%.

A figura 17 apresenta as variáveis significativas para o modelo II de regressão logística, esclarecendo as categorias apresentadas: *estimate* apresenta valores estimados para cada parâmetro em escala *logit,* ou seja, calcula o logaritmo em razão de chances; *std.error* é o desvio padrão relativo à estimativa pontual do coeficiente; *"Pr(*>|*z*|*)"*, também conhecido como *valor-p*, e utilizado para apurar a estimativa pontual do coeficiente, é significativamente diferente de 0; *z value* apresenta o número de desvios-padrão em relação à média da população.

Figura 17 - Coeficientes das variáveis significativas para amostra 75%

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  4.306e+02  7.767e+01   5.544 2.95e-08 ***
x1          -3.628e-01  2.791e-02 -13.001  < 2e-16 ***
x2          -2.131e-01  3.843e-02  -5.544 2.95e-08 ***
x3          -2.689e-05  1.901e-05  -1.414 0.157297
x4          -7.250e-01  1.562e-01  -4.642 3.46e-06 ***
x5           1.025e-02  1.443e-01   0.071 0.943405
x6          -2.710e-03  5.043e-03  -0.537 0.591023
x7          -7.428e-03  7.028e-03  -1.057 0.290561
x8           3.655e-02  7.606e-02   0.480 0.630896
x9           3.710e-02  2.501e-02   1.484 0.137925
x10          8.209e-02  1.352e-02   6.072 1.26e-09 ***
x12          3.827e-03  2.194e-02   0.174 0.861559
x13          1.007e-01  2.816e-02   3.577 0.000347 ***
x14         -5.665e-02  3.114e-02  -1.819 0.068880 .
x15          7.521e-02  7.003e-02   1.074 0.282871
x16         -9.835e-07  9.699e-07  -1.014 0.310587
x17         -1.061e-03  6.285e-04  -1.688 0.091319 .
x18          7.008e-03  6.522e-02   0.107 0.914432
x19          3.039e-04  3.753e-04   0.810 0.418083
x20         -5.534e-02  3.401e-02  -1.627 0.103707
x21         -1.692e-04  1.660e-04  -1.019 0.308019
x23          3.240e-01  5.593e-02   5.793 6.90e-09 ***
x24          5.533e-01  2.197e-01   2.518 0.011788 *
x25         -5.469e-02  2.192e-02  -2.495 0.012578 *
x26         -2.116e-03  1.857e-03  -1.139 0.254523
x27          1.150e-02  2.300e-02   0.500 0.617066
x28          1.018e-03  3.266e-04   3.117 0.001828 **
x29         -5.471e-03  3.445e-03  -1.588 0.112262
x30          4.304e-03  7.817e-03   0.551 0.581914
x31         -1.369e-05  3.008e-06  -4.551 5.33e-06 ***
x32          7.535e-02  5.376e-02   1.402 0.160975
x33          1.192e-03  7.375e-04   1.616 0.106068
x34         -9.271e-05  3.257e-04  -0.285 0.775951
x35         -5.224e-02  2.565e-02  -2.037 0.041691 *
x39          9.929e-05  2.351e-05   4.223 2.41e-05 ***
x40          1.896e-04  1.159e-04   1.636 0.101941
x41          2.679e-03  9.243e-04   2.899 0.003748 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Fonte: Autoria própria (2022)

"Signif codes", que são as áreas destacadas em vermelho na figura 17, expressam os códigos de significância, demonstrando as variáveis significativas do ponto de vista estatístico, sendo [0, 0,001] com código de significância ***, [0,001, 0,01] **, [0,01, 0,05] * e [0,1, 1] não possuem código de significância. Com as variáveis significativas encontradas, foi desenvolvido o modelo preditivo regressão logística para a amostra de 75%, conforme apresentado abaixo.

***Script* 3 - Modelo II de regressão logística**

```
> modeloII<-glm(y~x1+x2+x4+x10+x11+x13+x14+x17+x23+x24+x25+x28+x31
+ +x35+x39+x40+x41,family=binomial(link="logit"))
```

Fonte: Autoria própria (2022)

Para a amostra de 25% selecionada de maneira aleatória por meio da validação cruzada, as variáveis significativas são identificadas, permitindo o desenvolvimento do modelo preditivo III de regressão logística. Os resultados da identificação das variáveis significativas são expostos na figura 18.

Figura 18 - Coeficientes das variáveis significativas para amostra 25%

```
Coefficients: (4 not defined because of singularities)
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  4.090e+02  1.346e+02   3.039  0.00238 **
x1          -2.963e-01  4.646e-02  -6.377  1.8e-10 ***
x2          -3.090e-01  1.851e-01  -1.670  0.09496 .
x3          -2.042e-05  3.281e-05  -0.622  0.53373
x4          -9.898e-01  2.687e-01  -3.684  0.00023 ***
x5           2.106e-02  2.399e-01   0.088  0.93004
x7           1.738e-03  1.257e-02   0.138  0.89005
x8           5.633e-02  1.296e-01   0.435  0.66371
x9           5.773e-02  4.469e-02   1.292  0.19651
x10          7.047e-02  2.246e-02   3.138  0.00170 **
x12          5.511e-02  3.711e-02   1.485  0.13759
x13          6.406e-02  4.530e-02   1.414  0.15730
x14         -4.421e-02  5.288e-02  -0.836  0.40308
x16         -1.783e-07  1.606e-07  -1.110  0.26687
x17         -7.669e-04  1.349e-03  -0.569  0.56964
x18          4.242e-02  1.084e-01   0.391  0.69546
x19          8.999e-04  6.845e-04   1.315  0.18864
x20          1.150e-02  6.003e-02   0.192  0.84813
x21          1.050e-01  1.817e-01   0.578  0.56345
x22                 NA         NA      NA       NA
x23          1.756e-01  9.245e-02   1.900  0.05748 .
x24          3.954e-01  3.564e-01   1.109  0.26737
x25         -3.899e-02  3.562e-02  -1.095  0.27369
x26          6.594e-03  3.549e-03   1.858  0.06319 .
x28          1.743e-03  5.940e-04   2.935  0.00334 **
x29         -8.290e-03  6.178e-03  -1.342  0.17968
x30          2.369e-03  1.969e-02   0.120  0.90421
x31         -1.608e-05  5.616e-06  -2.862  0.00420 **
x32         -3.627e-02  9.918e-02  -0.366  0.71458
x33          2.013e-03  1.199e-03   1.679  0.09316 .
x35         -7.888e-02  4.404e-02  -1.791  0.07329 .
x36                 NA         NA      NA       NA
x37                 NA         NA      NA       NA
x38                 NA         NA      NA       NA
x39          1.365e-04  4.182e-05   3.265  0.00109 **
x40          1.659e-04  2.430e-04   0.683  0.49479
x41          2.579e-03  1.478e-03   1.744  0.08108 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Fonte: Autoria própria (2022)

As regras para encontro das significâncias da amostra são explicitadas na amostra de 75%; desta forma, para a amostra de 25%, temos o modelo preditivo III de regressão logística descrito abaixo:

***Script* 4 - Modelo III de regressão logística**

```
> modeloIII<-glm(y~x1+x4+x10+x11+x26+x31+x35+x39,
+ family=binomial(link="logit"))
```

Fonte: Autoria própria (2022)

O desvio nulo (2320 graus de liberdade) próximo do desvio residual (2303), para amostra de 75% da população e amostra de 25% desvio nulo (773 graus de liberdade) próximo do desvio residual (765), o modelo apenas com as variáveis significantes da amostra 75%, os graus de liberdade estão mais próximos, e o AIC (1523) é menor. Já na amostra de 25%, os graus de liberdade estão mais próximos, e o AIC (603,07) é maior. A tabela 3 apresenta os ajustes de desvio, grau de liberdade e a AIC do modelo aplicado para cada conjunto. Já a tabela 4 destaca a estatística descritiva dos modelos.

Tabela 4 - Ajuste dos modelos segundo as amostras

		Desvio	Graus de Liberdade	AIC
75%	Desvio nulo	2302,8	2320	1523
	Desvio residual	1487,0	2303	
25%	Desvio nulo	794,14	773	603,07
	Desvio residual	585,07	765	

Fonte: Autoria própria (2022)

O teste de *Hosmer-Lemeshow* para amostra de 25% (com todas as variáveis do modelo) apresentou os seguintes resultados: X^2=13,127, df = 8 e um p-valor = 0,1076; como o p-valor acima de 0,05. Os dados do modelo de regressão logística estão ajustados segundo do modelo de regressão logística. O teste de *Hosmer-Lemeshow* (apenas com as variáveis significativas do modelo anterior) X^2=38,587, df = 8 e um p-valor =5,86x10^ (-06); como o p-valor abaixo de 0,05. Os dados do modelo de regressão logística não estão ajustados conforme o modelo de regressão logística.

Tabela 5 - Estatística descritiva dos modelos

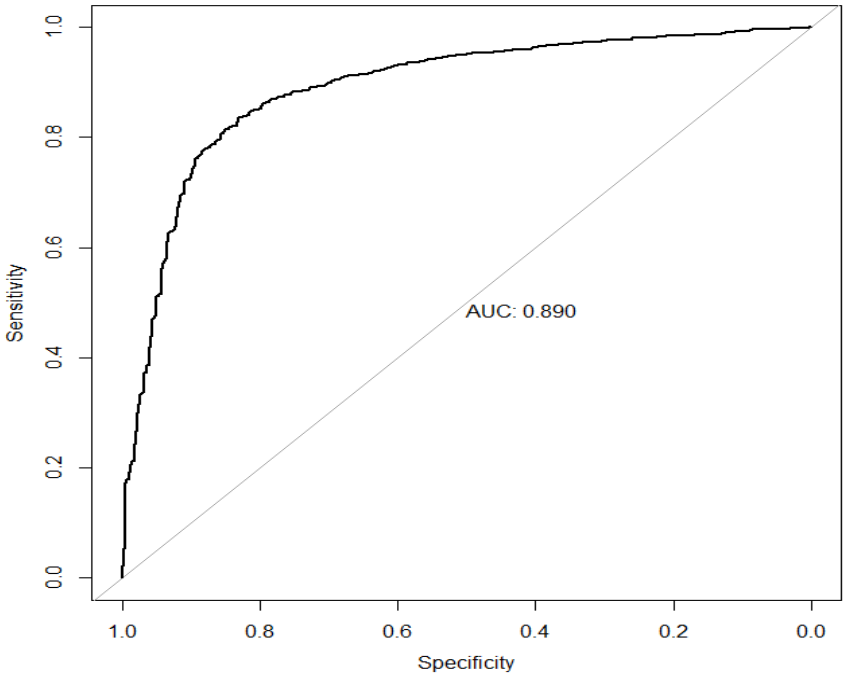
	Mínimo	1Q	Mediana	3Q	Máximo
75%	-31.052	0.1490	0.2833	0.4688	24.525
25%	-25.463	0.2821	0.3657	0.5283	19.147

Fonte: Autoria própria (2022)

A validação interna do modelo II ocorreu por meio da curva de características operacionais do receptor (ROC), uma representação gráfica que apresenta o

desempenho do modelo, contribuindo à tomada de decisão sobre o melhor valor-limite e para identificar seu poder preditivo. Um valor de limiar alto fornece alta especificidade e baixa sensibilidade; um valor de limiar baixo fornece baixa especificidade e alta sensibilidade. O gráfico 10 destaca como a curva ROC e a área AUC ficam dispostas para o conjunto de treinamento 75%.

Gráfico 9 - Curva de características do receptor (ROC) para validação do modelo II do conjunto de treinamento 75%



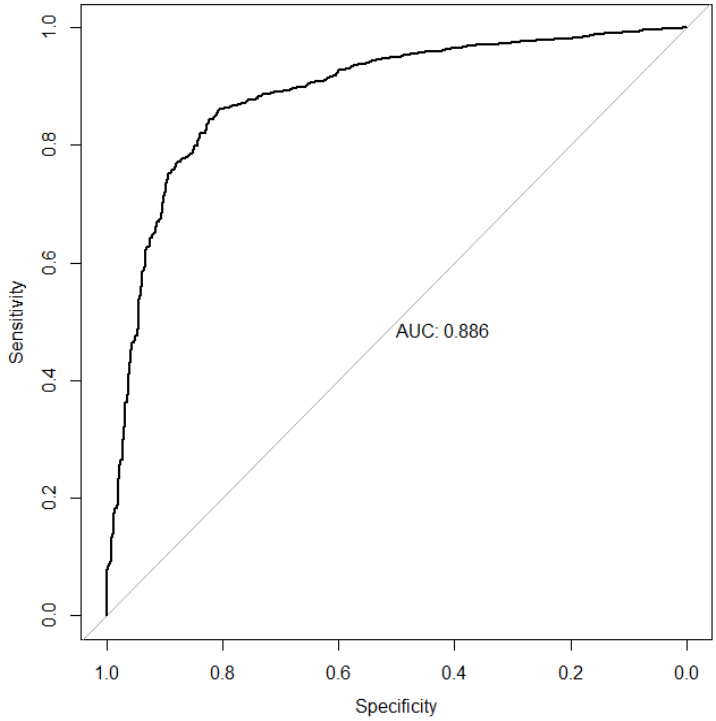
Fonte: Autoria própria (2022)

Apresentando uma análise comparativa entre as taxas verdadeiro positivo, que exploram a sensibilidade, e a taxa de falso positivo, a qual remete à especificidade do gráfico 10 da ROC, a curva de sensibilidade apresenta os pontos de corte na probabilidade estimada, conhecida como "*thresholds*". Para o conjunto de 75%, demonstra algumas variações entre a taxa verdadeiro positivo e a taxa falso positivo. Isto demonstra que o modelo é casual, pelo comportamento dos *theresholds,* iniciando com os valores preditos, não considerados TVP ou TFP. Ao final, em 0% conclui-se que todos os TVPs foram preditos como 1, e a TFP atinge 100%.

A área do gráfico que fica sob a curva (AUC) expõe valores por meio da medida da área, expressados por 0,0 a 1,0 ou 0% a 100%, e consenso entre autores que um alto valor pode representar um desempenho aceitável do modelo preditivo

desenvolvido (JOHNSTON *et al.,* 2019; HAQ *et al.,* 2020). No gráfico 10, a AUC tem valor 0,890, destacando que um percentual de precisões está correto.

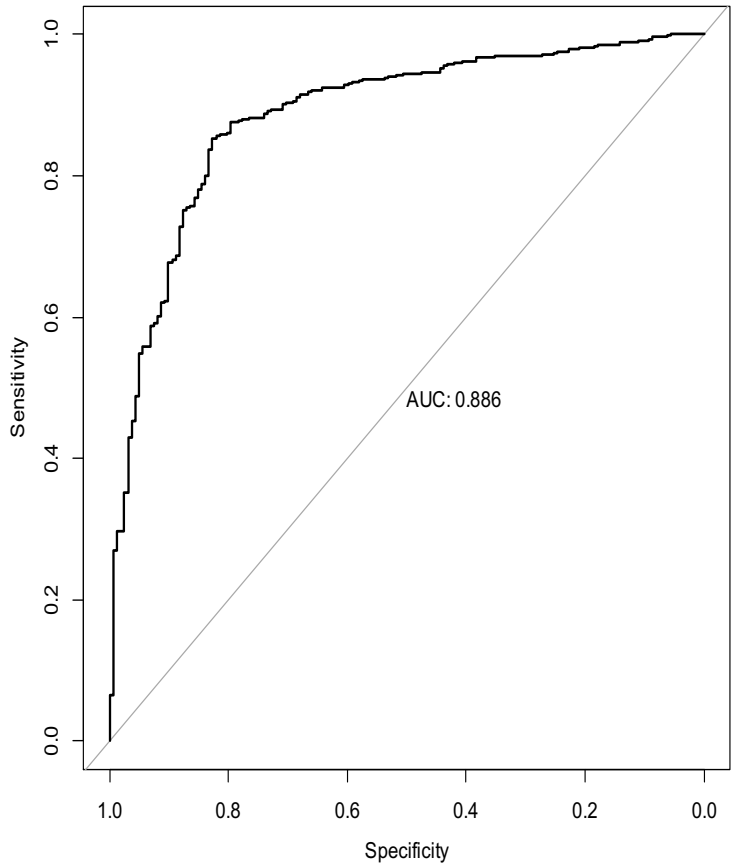
Gráfico 10 - Curva de características do receptor (ROC) para validação do modelo II do conjunto de teste 25%



Fonte: Autoria própria (2022)

No gráfico 11, a curva ROC apresenta variações sobre os *thresholds*, para o modelo preditivo do conjunto de teste 25%. O comportamento apurado destaca que as curvas ROC e AUC apresentam algumas semelhanças. Para superar esta questão, a medicação da curva AUC expressa em termos numéricos um valor diferente, ou seja, AUC para o conjunto de 25% é 0.886. Ainda assim, este resultado destaca o desempenho aceitável.

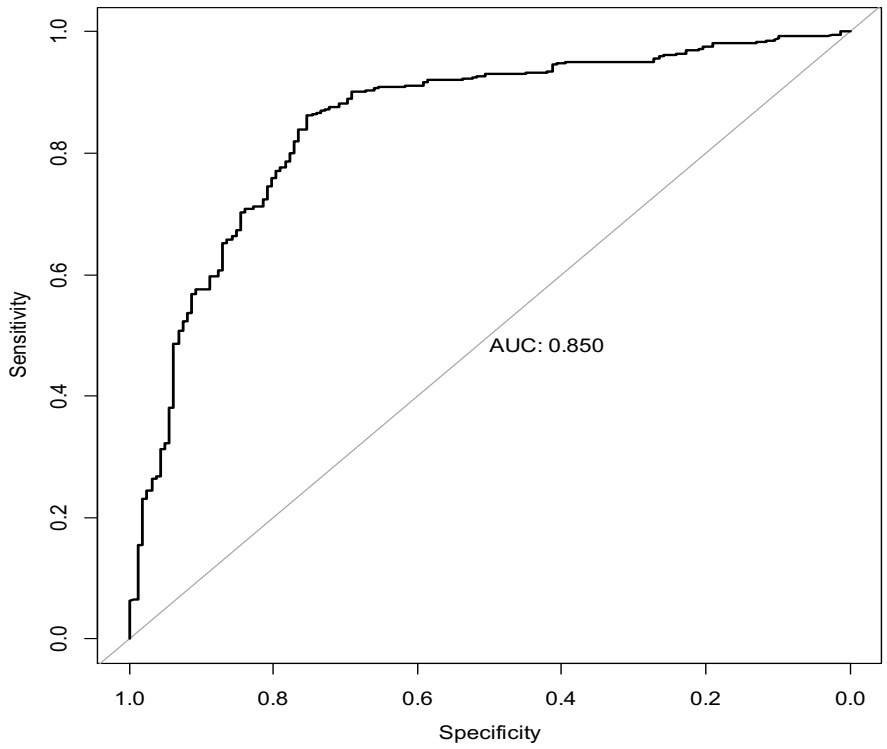
Gráfico 11 - Curva de características do receptor (ROC) para validação do modelo III do conjunto de treinamento 75%



Fonte: Autoria própria (2022)

No gráfico 12, o modelo III foi aplicado ao conjunto de dados de treinamento de 75%; a variação dos *thresholds* é visivelmente variável; a curva AUC expõe o valor 0.886, o qual destaca que as predições estão corretas diante do comportamento das taxas de TVP e TFP.

Gráfico 12 - Curva de características do receptor (ROC) para validação do modelo III do conjunto de teste 25%



Fonte: Autoria própria (2022)

Conforme observado no gráfico 13, o comportamento das taxas de TVP e TFP possui uma variação, medição da curva AUC a qual apresenta um valor de 0.850 que permite compreender que o desempenho do modelo preditivo é aceitável diante das variações existentes.

4.2.2 Algoritmo árvore de decisão aplicado ao *dataset bd_oncology*

O algoritmo árvore de decisão foi utilizado para ajustar a tomada de decisão, representada por meio de um gráfico no formato de árvore, esclarecendo as probabilidades para se chegar a resultados. Para construir a árvore, utilizou-se a função *rpart* do programa *R*, versão 4.0.3. O *script* desenvolvido é apresentado inicialmente.

Script* 5 - Árvore de decisão aplicada ao *bd_oncology* no *software R

```
rm(list = ls())
#install.packages("plyr")
```

```
#install.packages("readr")
#install.packages("dplyr")
#install.packages("caret")
#install.packages("rpart.plot")
library(tcltk) # TCL/TK para abrir o bd
library(rpart)# Arvore de decisão
library(car)
library(MASS)
library(plyr)
library(readr)
library(dplyr)
library(caret)
library(rpart)
library(rpart.plot)
BD<-read.csv2("E:/MYLLER/BANCOS/BD75.csv",header=T)
attach(BD)
names(BD)
str(BD)
#VARIAVEIS#
y<-BD$retornou.ou.não
x1<-BD$diagant
x2<-BD$ANO
x3<-BD$recno
x4<-BD$tpcaso
x5<-BD$sexo
x6<-BD$idade
x7<-BD$local777s
x8<-BD$racacor
x9<-BD$instruc
x10<-BD$cliaten
x11<-BD$clitrat
x12<-BD$histfamc
x13<-BD$alcoolis
x14<-BD$tabagism
x15<-BD$estadres
x16<-BD$proceden
x17<-BD$ano18idi
x18<-BD$orienc
x19<-BD$exdiag
x20<-BD$estconj
x21<-BD$antri
x22<-BD$dt18icon
x23<-BD$basmaimp
x24<-BD$Loctudet
x25<-BD$Loctu18i
x26<-BD$tipohist
x27<-BD$laterali
x28<-BD$tnm
```

```
x29<-BD$estadiam
x30<-BD$outroesta
x31<-BD$ptnm
x32<-BD$rzntr
x33<-BD$dtinitrt
x34<-BD$A18itrath
x35<-BD$estdfimt
x36<-BD$cnes
x37<-BD$ufuh
x38<-BD$muuh
x39<-BD$ocupacao
x40<-BD$d_tri_diag
x41<-BD$d_1Cons_Trat
table(y)
glimpse(BD)
modelo<-rpart(y~x1+x2+x3+x4+x5+x6+x7+x8+x9+x10+x11+x12+x13+x14+x15+x16+x17+x18
+x19+x20+x21+x22+x23+x24+x25+x26+x27+x28+x29+x30+x31+x32+x33+x34+x35+x36+x37
+x38+x39+x40+x41,data = BD, method="class", minsplit = 10, minbucket=3)
summary(modelo)
prp(modelo,
faclen=0, #use full names for factor labels
extra=1, #display number of obs. for each terminal node
roundint=F, #don't round to integers in output
digits=5) #display 5 decimal places in output
predict(modelo, type = "prob")   # probabilidades de classe (padrão)
predict(modelo, type = "vector") # números de nível
predict(modelo, type = "class")  # fator
predict(modelo, type = "matrix")
PredictCART_train = predict(modelo, data = BD, type = "class")
table(PredictCART_train)#
#(2536+82)/(2536 + 82 + 477)
rpart.plot(modelo,type = 3)
```

Fonte: Autoria própria (2022)

Modelo para árvore de decisão aplicado ao conjunto de dados 75%; o *script* abaixo destaca como ele foi formatado.

***Script* 6 - Modelo de árvore de decisão aplicado ao conjunto de treinamento 75%**

```
Call:
rpart(formula = y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + x9 +
    x10 + x11 + x12 + x13 + x14 + x15 + x16 + x17 + x18 + x19 +
    x20 + x21 + x22 + x23 + x24 + x25 + x26 + x27 + x28 + x29 +
    x30 + x31 + x32 + x33 + x34 + x35 + x36 + x37 + x38 + x39 +
    x40 + x41, data = BD, method = "class", minsplit = 10, minbucket = 3)
  n= 2321
```

Fonte: Autoria própria (2022)

Os resultados da árvore de decisão apurados pela função *rpart*, função esta que possui um procedimento de validação cruzada, o qual desenvolve os conjuntos de dados para treinamento e teste. Os resultados estão dispostos aqui em parâmetros.

Parâmetro CP - *"complexity parameter"*: é utilizado para controlar o tamanho da árvore, e corresponde ao incremento de menor custo do modelo necessário para adicionar uma variável nova. Parâmetro "*rel error*": estabelece o erro relativo à classificação dos dados do conjunto de treinamento obtidos por meio da árvore. Parâmetro *"xerror"*: mede o erro relativo à classificação dos dados do conjunto de teste, efetuado por meio da árvore de decisão construída com o conjunto de treinamento. Parâmetro *"xstd"*: é o qual representa o erro padrão. Esses resultados são apresentados na tabela 6 a seguir.

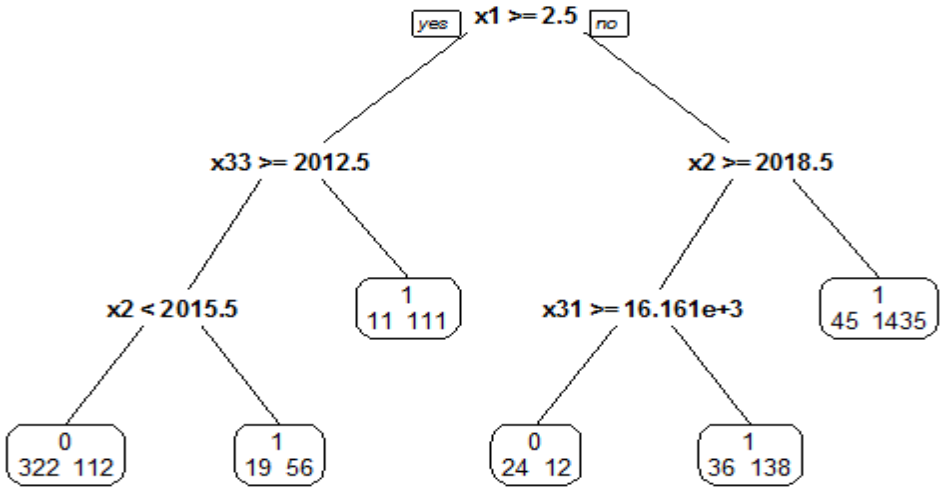
Tabela 6 - Resultados da função *rpart* para poda da árvore

	CP	rel	Error	Xerror	Xstd
1	0,189278	0	1,000000	1,000000	0,041921
2	0,080963	2	0,621444	0,632385	0,034806
3	0,013129	3	0,540481	0,549234	0,032739
4	0,010000	5	0,514223	0,551422	0,032796

Fonte: Autoria própria (2022)

A coluna "*xerror*" contém estimativas de erro, sendo significativa a previsão com validação cruzada para diferentes números de divisões (*nsplit*). As variáveis importantes para o modelo de treinamento 75%: x1, x33, x2, x22, x21, x17, x3, x39, x19, x31 x41 x10 x23. A figura 19 demonstra como a árvore de decisão se configurou para o conjunto de treinamento de 75%.

Figura 19 - Diagrama árvore de decisão para o conjunto de treinamento 75% - função (prp)



Fonte: Autoria própria (2022)

A função *"summary"* no *R* desenvolve apuração dos resultados e, consequentemente, destaca em cada detalhe do modelo previsto o grau de importância de cada variável e as proporções para cada classe. Na sequência, são apresentadas as probabilidades de cada nó de decisão.

O preditor x1 (variável de diagnóstico e tratamento anteriores dos pacientes) desenvolve a primeira divisão de dados considerada o nó-raiz; possui 2.321 observações, correlaciona com o preditor x33 (positivo) com probabilidade igual 0,197 e com o preditor x2 (negativo) com probabilidade do nó de 0.803, ou seja, os pacientes que não têm histórico de câncer.

No preditor x33 (ano do início do primeiro tratamento específico para o tumor, no hospital), para os pacientes que não têm histórico de câncer, há uma probabilidade do nó de 0.272. Para o preditor do lado direito, apresentam 122 observações com a probabilidade do nó igual a 0,442; e do lado esquerdo, apresentam 509 observações com a probabilidade do nó igual a 0,558.

No preditor x2 (ano do diagnóstico), para os pacientes que não têm histórico de câncer, apresenta-se um preditor do lado esquerdo com uma contagem de classes [322; 112] com umas probabilidades de [0.742;0.258] respectivamente; e do lado direito, com uma contagem de classes [19;56] com umas probabilidades de [0.253;0.747] respectivamente.

No preditor x2 (ano do diagnóstico), para os pacientes que têm histórico de câncer, apresentam-se dois preditores; para o lado direito, uma contagem de classes

[45; 1435] com uma probabilidade de [0.030;0.970]; e do lado esquerdo, o preditor x2 está ligado ao preditor x31 (classificação ptnm) que apresenta duas ramificações; do lado esquerdo, são 36 pacientes que não têm câncer, apresentando uma contagem de classes [24; 12] com umas probabilidades iguais a [0.667;0.333]; e do lado direito, apresenta-se uma contagem de classes [36;138] com umas probabilidades iguais a [0.207; 0.793].

A perda esperada expõe a soma dos valores de todas as perdas possíveis; cada uma multiplicada pela probabilidade de essa perda ocorrer. No conjunto de teste desenvolvido por meio da função *rpart,* um modelo de árvore de decisão, considerando todas as variáveis, é desenvolvido. O *script* abaixo destaca sua notação.

***Script* 7 - Modelo de árvore de decisão aplicado ao conjunto de teste 25%**

```
Call:
rpart(formula = y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + x9 +
    x10 + x11 + x12 + x13 + x14 + x15 + x16 + x17 + x18 + x19 +
    x20 + x21 + x22 + x23 + x24 + x25 + x26 + x27 + x28 + x29 +
    x30 + x31 + x32 + x33 + x34 + x35 + x36 + x37 + x38 + x39 +
    x40 + x41, data = BD, method = "class", minsplit = 10, minbucket = 3)
  n= 774
```

Fonte: Autoria própria (2022)

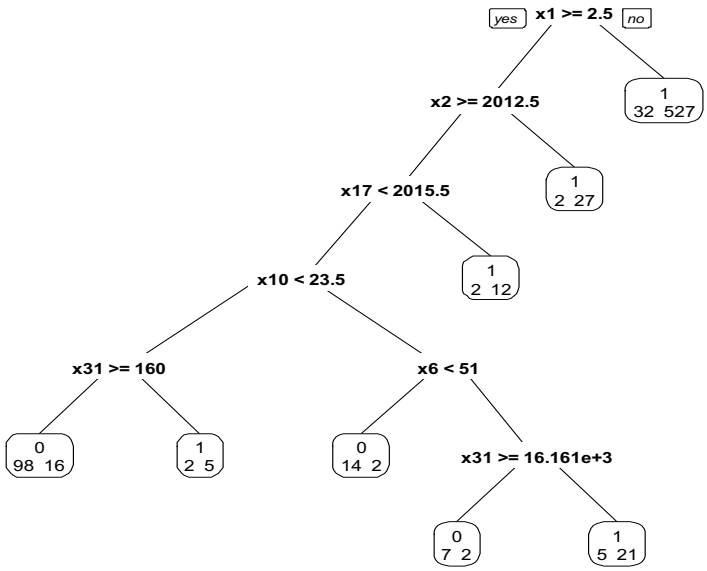
A tabela 7 a seguir destaca os resultados encontrados no experimento do modelo preditivo árvore de decisão aplicado ao conjunto de teste 25%; este conjunto é determinado pela função *rpart* a qual possui a validação cruzada de maneira integrada. A coluna *xerror* contém estimativas de erro as quais são significativas à previsão com validação cruzada para diferentes números de divisões *"nsplit"*. As variáveis importantes para o modelo de teste são: x1, x2, x22, x21, x17, x33, x31, x10, x6, x19, x3, x41, x39, x11, x4, x23.

Tabela 7 - Resultados da função *rpart* para poda da árvore conjunto de teste 25%

	CP	rel	Error	Xerror	Xstd
1	0.27777778	0	1.0000000	1.0000000	0.06986308
2	0.15432099	1	0.7222222	0.7222222	0.06151629
3	0.06172840	2	0.5679012	0.5679012	0.05557770
4	0.03395062	3	0.5061728	0.5555556	0.05505082
5	0.03086420	5	0.4382716	0.5432099	0.05451524
6	0.01851852	6	0.4074074	0.5123457	0.05313645
7	0.01000000	7	0.3888889	0.5061728	0.05285359

Fonte: Autoria própria (2022)

Figura 20 - Diagrama árvore de decisão para o conjunto de teste 25% - função (prp)



Fonte: Autoria própria (2022)

Na árvore de decisão para amostra de 25%, o preditor x1 (variável de diagnóstico e tratamento anteriores dos pacientes) tem duas ramificações: o lado direito apresenta 559 pacientes que não possuem histórico de câncer, com uma probabilidade do nó igual a 0.722, contagens de classe de [32; 527] e probabilidades de [0.057;0.943] respectivamente; e o lado esquerdo, uma correlação com o preditor x2.

O preditor x2, do lado esquerdo tem o preditor x17 e do lado direito tem uma probabilidade do nó de 0.037; contagens de classe [2; 27] e probabilidade de [0.069; 0.931].

O preditor x17, do lado esquerdo tem o preditor x10 e do lado direito tem uma probabilidade do nó igual a 0.018; contagens de classe [2; 12] e probabilidade de [0.143; 0,857].

O preditor x10: do lado esquerdo, tem o preditor x31; e do lado direito, o preditor x6. O preditor x31 apresenta uma probabilidade do nó igual a 0.147; contagens de classe [98; 16] e uma probabilidade [0.860;0.140]; do lado esquerdo, a probabilidade do nó é igual 0.009; contagens de classe [2; 5] e a probabilidade de [0.286; 0.714].

O preditor x6 do lado direito o preditor x31 do lado esquerdo, apresentam uma probabilidade do nó de 0.021; contagens de classe [14; 2]; com uma probabilidade [0.875;0.125]; o preditor x31 apresenta duas ramificações; do lado esquerdo, a

probabilidade do nó é igual 0.012; contagens de classe [7; 2], e as probabilidades de [0.778; 0.222]; e do lado direito, apresenta uma probabilidade do nó de 0.033; contagens de classe [5; 21], e as probabilidade de [0.192;0.808].

4.2.3 Testes de diagnóstico para validação de desempenho árvore de decisão

Orientado pelo estágio 3 do modelo preditivo previsto na metodologia, os testes de diagnósticos produzidos foram aplicados ao algoritmo árvore de decisão desenvolvido para todos os conjuntos de dados, ou seja, treinamento e teste. Conjunto de treinamento 75% foi o primeiro a ser testado, o *script* a seguir destaca as medidas de desempenho e teste de diagnóstico desenvolvido:

***Script* 8 - Medidas de desempenho e teste de diagnóstico aplicado na árvore de decisão conjunto de treinamento e teste**

```
rm(list = ls())
library(tcltk) # TCL/TK para abrir o bd
BD<-read.csv2("E:/MYLLER/BANCOS/BD2019.csv",header=T)
attach(BD)
names(BD)
#VARIAVEIS#
x1<-BD$retornou1
x2<-BD$tpcaso
x3<-BD$sexo
####TABELAS CRUZADAS (2X2) #####
T1<-table(x1,x3)
T2<-table(x1,x2)
T3<-table(x2,x3)
#### RETORNOU OU NÃO X SEXO ######
a1<-T1[1,1]
b1<-T1[1,2]
c1<-T1[2,1]
d1<-T1[2,2]
#Indice de Kappa
Po1<-(a1+d1)/a1+b1+c1+d1
Pe1<-((a1 + b1)*(a1 + c1)) + ((c1 + d1)*(b1 + d1))/(a1+b1+c1+d1)^2
K1<-(Po1-Pe1)/(1-Pe1)
#Sensibilidade:
```

```
s1<-(a1/(a1+c1))
#Especificidade:
e1<-(d1/(b1+d1))
#Prevalência (real):
P1<-(a1+c1)/(a1+b1+c1+d1)
#Prevalência estimada (teste):
PE1<-(a1+b1)/(a1+b1+c1+d1)
#Valor preditivo positivo:
VPP1<-(a1)/(a1+b1)
#Valor preditivo negativo:
VPN1<-(d1)/(c1+d1)
#Classificação correta (acurácia):
ac1<-(a1+d1)/(a1+b1+c1+d1)
#Classificação incorreta:
CI1<-(b1+c1)/(a1+b1+c1+d1)
#### RETORNOU OU NÃO X TP CASO ####
a2<-T2[1,1]
b2<-T2[1,2]
c2<-T2[2,1]
d2<-T2[2,2]
#Indice de Kappa
Po2<-(a2+d2)/a2+b2+c2+d2
Pe2<-((a2 + b2)*(a2 + c2)) + ((c2 + d2)*(b2 + d2))/(a2+b2+c2+d2)^2
K2<-(Po2-Pe2)/(2-Pe2)
#Sensibilidade:
s2<-(a2/(a2+c2))
#Especificidade:
e2<-(d2/(b2+d2))
#Prevalência (real):
P2<-(a2+c2)/(a2+b2+c2+d2)
#Prevalência estimada (teste):
PE2<-(a2+b2)/(a2+b2+c2+d2)
#Valor preditivo positivo:
VPP2<-(a2)/(a2+b2)
#Valor preditivo negativo:
```

```
VPN2<-(d2)/(c2+d2)
#Classificação correta (acurácia):
ac2<-(a2+d2)/(a2+b2+c2+d2)
#Classificação incorreta:
CI2<-(b2+c2)/(a2+b2+c2+d2)
#### TP SEXO X SEXO ####
a3<-T2[1,1]
b3<-T2[1,2]
c3<-T2[2,1]
d3<-T2[2,2]
#Indice de Kappa
Po3<-(a3+d3)/a3+b3+c3+d3
Pe3<-((a3 + b3)*(a3 + c3)) + ((c3 + d3)*(b3 + d3))/(a3+b3+c3+d3)^3
K3<-(Po3-Pe3)/(3-Pe3)
#Sensibilidade:
s3<-(a3/(a3+c3))
#Especificidade:
e3<-(d3/(b3+d3))
#Prevalência (real):
P3<-(a3+c3)/(a3+b3+c3+d3)
#Prevalência estimada (teste):
PE3<-(a3+b3)/(a3+b3+c3+d3)
#Valor preditivo positivo:
VPP3<-(a3)/(a3+b3)
#Valor preditivo negativo:
VPN3<-(d3)/(c3+d3)
#Classificação correta (acurácia):
ac3<-(a3+d3)/(a3+b3+c3+d3)
#Classificação incorreta:
CI3<-(b3+c3)/(a3+b3+c3+d3)
############################
#######RESULTADOS##########
############################
#### RETORNOU OU NÃO X SEXO ######
K1 #Indice de Kappa
```

```
s1 #Especificidade:
e1 #Especificidade:
P1 #Prevalência (real):
PE1  #Prevalência estimada (teste):
VPP1 #Valor preditivo positivo:
VPN1 #Valor preditivo negativo:
ac1  #Classificação correta (acurácia):
CI1  #Classificação incorreta:
#### RETORNOU OU NÃO X TP CASO ####
K2 #Indice de Kappa
s2 #Especificidade:
e2 #Especificidade:
P2 #Prevalência (real):
PE2  #Prevalência estimada (teste):
VPP2 #Valor preditivo positivo:
VPN2 #Valor preditivo negativo:
ac2  #Classificação correta (acurácia):
CI2  #Classificação incorreta:
#### TP SEXO X SEXO ####
K3 #Indice de Kappa
s3 #Especificidade:
e3 #Especificidade:
P3 #Prevalência (real):
PE3  #Prevalência estimada (teste):
VPP3 #Valor preditivo positivo:
VPN3 #Valor preditivo negativo:
ac3  #Classificação correta (acurácia):
CI3  #Classificação incorreta:
```

Fonte: Autoria própria (2022)

Confeccionou-se a tabela 2x2, considerando a variável dependente, retornou ou não ao tratamento, e sua relação com a variável independente sexo do paciente. Os dados são apresentados na tabela a seguir.

Tabela 8 - Tabela 2x2 das variáveis retornou ou não ao tratamento e sexo do paciente

Tratamento	Masculino	Feminino	Total
Sim	848	1.016	1.864
Não	217	240	457
Total	1.065	1.256	2.321

Fonte: Autoria própria (2022)

Na sequência, são descritos todos os resultados das medidas de desempenho definidas pelo estágio 3 do modelo preditivo; a sua intenção é garantir a capacidade preditiva do modelo produzido e seus aspectos que reportam sua qualidade.

1. Índice de Kappa: 0.9992579 (Ótima)
2. Sensibilidade: 0.7962441
3. Especificidade: 0.1910828
4. Prevalência (real): 0.4588539
5. Prevalência estimada (teste): 0.8031021
6. Valor preditivo positivo: 0.4549356
7. Valor preditivo negativo: 0.5251641
8. Classificação correta (acurácia): 0.4687635
9. Classificação incorreta: 0.5312365

Na tabela 2x2, a variável “retornou ou não ao tratamento” correlaciona a variável “tipo de caso do paciente”, se correlacionam. A variável “tipo de caso do paciente” pode ser classificada um caso analítico e não analítico. Os casos analíticos são novas ocorrências de câncer, possuindo um planejamento terapêutico produzido pelo hospital. Os casos não analíticos se referem aos casos que já desenvolveram algum tratamento, sendo admitidos para um novo ou complementação do tratamento recomendado. A tabela 9 destaca resultados dos testes de diagnósticos e medidas de desempenho.

Tabela 9 - Tabela 2x2 das variáveis retornou ou não ao tratamento e tipo de caso do paciente

Tratamento	Analíticos	Não analíticos	Total
Sim	1.299	565	1.864
Não	149	308	457
Total	1.448	873	2.321

Fonte: Autoria própria (2022)

Neste sentido, são descritos todos os resultados das medidas de desempenho definidas pelo estágio 3 do modelo preditivo.

1. Índice de Kappa: 0.9996216 (Ótima)
2. Sensibilidade: 0.8970994
3. Especificidade: 0.3528064
4. Prevalência (real): 0.623869
5. Prevalência estimada (teste): 0.8031021
6. Valor preditivo positivo: 0.6968884
7. Valor preditivo negativo: 0.6739606
8. Classificação correta (acurácia): 0.692374
9. Classificação incorreta: 0.307626

Os resultados das correlações das variáveis sexo em relação ao tipo de caso do paciente estão dispostos na tabela a seguir.

Tabela 10 - Tabela 2x2 das variáveis sexo e tipo de caso do paciente

Sexo	Analíticos	Não analíticos	Total
Masculino	646	802	1.448
Feminino	419	454	873
Total	1.065	1.256	2.321

Fonte: Autoria própria (2022)

Na sequência, são apresentados os resultados dos testes e medidas de desempenho aplicados ao modelo árvore de decisão sobre o conjunto de treinamento 75%.

1. Índice de Kappa: 0.999622 (Ótima)
2. Sensibilidade: 0.8970994
3. Especificidade: 0.3528064
4. Prevalência (real): 0.623869
5. Prevalência estimada (teste): 0.8031021
6. Valor preditivo positivo: 0.6968884
7. Valor preditivo negativo: 0.6739606
8. Classificação correta (acurácia): 0.692374
9. Classificação incorreta: 0.307626

Com a mesma intencionalidade, as medidas de desempenho são aplicadas ao conjunto de dados de teste 25%; sua verificação ocorre com as mesmas variáveis analisadas no conjunto de treinamento 75%, conjuntamente; os resultados das medidas de desempenho são descritos após a apresentação da tabela 2x2, assim exposta:

Tabela 11 - Tabela 2x2 das variáveis retornou ou não ao tratamento e sexo do paciente.

Tratamento	Masculino	Feminino	Total
Sim	276	336	612
Não	78	84	162
Total	354	420	774

Fonte: Autoria própria (2022)

Os resultados das medidas de desempenho são destacados a seguir.

1. Índice de Kappa: 0.9976999 (Ótima)
2. Sensibilidade: 0.779661
3. Especificidade: 0.2
4. Prevalência (real): 0.4573643
5. Prevalência estimada (teste): 0.7906977
6. Valor preditivo positivo: 0.4509804
7. Valor preditivo negativo: 0.5185185
8. Classificação correta (acurácia): 0.4651163
9. Classificação incorreta: 0.5348837

Acerca das correlações entre as variáveis retornou ou não ao tratamento e tipo de caso do paciente, os resultados estão dispostos na tabela 2x2, na sequência.

Tabela 12 - Tabela 2x2 das variáveis retornou ou não ao tratamento e tipo de caso do paciente

Tratamento	Analíticos	Não analíticos	Total
Sim	431	181	612
Não	49	113	162
Total	480	294	774

Fonte: Autoria própria (2022)

Medidas de desempenho com seus respectivos resultados são exibidos na sequência.

1. Índice de Kappa: 0.9988349 (Ótima)
2. Sensibilidade: 0.8979167

3. Especificidade: 0.3843537
4. Prevalência (real): 0.620155
5. Prevalência estimada (teste): 0.7906977
6. Valor preditivo positivo: 0.7042484
7. Valor preditivo negativo: 0.6975309
8. Classificação correta (acurácia): 0.7028424
9. Classificação incorreta: 0.2971576

A tabela demonstrada a seguir destaca a correlação entre as variáveis sexo e tipo de caso do paciente. Os resultados estão assim dispostos:

Tabela 13 - Tabela 2x2 das variáveis sexo e tipo de caso do paciente

Sexo	Analíticos	Não analíticos	Total
Masculino	214	266	480
Feminino	140	154	294
Total	354	420	774

Fonte: Autoria própria (2022)

Resultados a seguir, destacam as medidas de desempenho, considerando as variáveis sexo e tipo de caso do paciente.

1. Índice de Kappa: 0.9988383 (Ótima)
2. Sensibilidade: 0.8979167
3. Especificidade: 0.3843537
4. Prevalência (real): 0.620155
5. Prevalência estimada (teste): 0.7906977
6. Valor preditivo positivo: 0.7042484
7. Valor preditivo negativo: 0.6975309
8. Classificação correta (acurácia): 0.7028424
9. Classificação incorreta: 0.2971576

Na próxima seção, são discutidos o modelo preditivo desenvolvido com duas possibilidades de predição, em comparação aos modelos encontrados na literatura.

4.3 Discussão sobre o modelo preditivo proposto

Modelos preditivos em cuidados de saúde são instrumentos capazes de apoiar o processo de tomada de decisão por meio de uma análise criteriosa sobre o imenso repertório de dados e informações que, muitas vezes, são desconsiderados em um processo de avaliação de diagnóstico e, consequentemente, após o primeiro tratamento. Com algumas funcionalidades de transferência de tecnologia, a literatura não apresenta muitas pesquisas envolvendo modelos preditivos aplicados a pacientes oncológicos utilizando algoritmos de aprendizagem supervisionada.

A proposta deste estudo foi desenvolver um modelo preditivo capaz de preditar resultados sobre pacientes submetidos a cuidados de saúde e diagnosticados com câncer. Para isso, foram utilizadas funcionalidades da análise *big data*, como comportamento dos dados, governança exercida, análise dos dados, agregação de dados. Além disso, contribuições do processo de transferência de tecnologia, como conhecimento novo produzido por meio das predições, *expertise* e habilidades objetivando a significância clínica.

Inicialmente foi construído um *dataset* com dados e informações disponibilizadas sobre os pacientes que foram diagnosticados com câncer e receberam o respectivo tratamento; dois algoritmos foram selecionados entendendo-se a sua capacidade de predição, regressão logística e árvore de decisão. Diante disso, o rótulo desenvolvido diante do problema observado foi acerca da probabilidade desses pacientes retornarem ou não ao serviço de oncologia, justamente por compreender-se que, após a conclusão do primeiro tratamento, existe um risco associado ao retorno ou não da doença aos mesmos locais ou localidades diversas.

Sendo assim, a principal intenção do modelo preditivo foi contribuir para o processo de gestão que envolve os serviços de oncologia do hospital investigado. Isso no sentido de compreender a possibilidade de evolução do quadro clínico por meio das predições. Além disso, apoiar as equipes clínicas, enfermagem e multiprofissionais, acerca do planejamento das intervenções desenvolvidas diariamente, procurando a eficiência e eficácia sobre a carga de trabalho.

Considerando a estrutura de preditores preconizada no RHC, entendendo-se sua abrangência nacional brasileira, se for aplicado em um outro contexto geográfico internacional, o ideal é compreender a aderência das variáveis dispostas no modelo,

e entender o modo com que elas são observadas pelos serviços de cuidados em saúde.

Na revisão de literatura, foram encontrados 17 modelos preditivos, demonstrados no quadro 7. Tais modelos foram aplicados a cuidados de saúde em diversos tipos de situações do processo de cuidados em saúde; cobrindo, assim, um escopo limitado de doenças, objetivando mais a predição e TT; especificamente sobre a doença do câncer não foram encontrados modelo com poder de predição voltado à ressubmissão de pacientes aos serviços de cuidados de saúde.

Quadro 7 - Modelos preditivos encontrados na revisão sistemática de literatura

Pesquisa	Ano	Foco do modelo preditivo
1. A predictive model of technology transfer using patent analysis	2015	Análise de patentes para o processo de TT
2. Big data analytics-based predictive modeling for stress management using healthcare system	2017	Foco na previsão do nível de estresse, considerando 4 atributos.
3. A machine-learning model for automatic detection of movement compensations in stroke patients	2020	Modelo voltado à reabilitação de pacientes de AVC.
4. Evaluation of machine learning methodology for the prediction of healthcare resource utilization and healthcare costs in patients with critical limb ischemia-is preventive and personalized approach on the horizon?	2020	Voltado à previsão personalizada do paciente por meio da linha de base antes do diagnóstico de isquemia crítica.
5. Automatic alarm prioritization by data mining for fault management in cellular networks	2020	Modelo com método automático de priorização de alarmes na gestão de falhas.
6. Heart Disease Identification Method Using Machine Learning Classification in E-Healthcare	2020	Sistema de predição de diagnóstico de doenças cardíacas.
7. Optimised Big Data analytics for health and safety hazards prediction in power infrastructure operations	2020	Técnica para encontrar padrões complexos, estabelecer coesão estatística em casos de acidentes em obras.
8. Healthcare pathway discovery and probabilistic machine learning	2020	Aplicação do aprendizado de máquina probabilístico para prever os tempos de recuperação específicos do paciente após a apendicectomia, usando informações de caminhos dos caminhos aprendidos e outros dados de EMR relevante.
9. Smart healthcare framework for ambient assisted living using IoMT and big data analytics techniques	2019	Estrutura de saúde inteligente para monitorar as atividades físicas dos idosos.
10. Towards a decision support tool for intensive care discharge: machine learning algorithm development using electronic healthcare data from MIMIC-III and Bristol, UK	2019	Método automatizado para detectar pacientes que estão prontos para alta em unidades de terapia intensiva.

11. Big data analytics and artificial intelligence pathway to operational performance under the effects of entrepreneurial orientation and environmental dynamism: A study of manufacturing organisations	2020	Modelo para descrever o papel da orientação empreendedora, baseado em capacidades dinâmicas.
12. Using Machine Learning Applied to Real-World Healthcare Data for Predictive Analytics: An Applied Example in Bariatric Surgery	2019	Modelo de cessação de medicação anti-hiperglicêmica em pacientes submetidos à cirurgia bariátrica.
13. Decision-Making based on Big Data Analytics for People Management in Healthcare Organizations	2019	Proposta de modelo preditivo aplicado à gestão de pessoas em organizações de saúde.
14. A technology delivery system for characterizing the supply side of technology emergence: Illustrated for Big Data & Analytics	2018	Modelo voltado ao sistema de entrega de tecnologias, caracterizado pelo lado da oferta do surgimento de tecnologia.
15. Intelligent Classifier: a Tool to Impel Drug Technology Transfer from Academia to Industry	2019	Modelo de classificador de máquina de vetor de suporte, usando os dados de patentes farmacêuticas mantidas por universidades para prever os resultados do licenciamento dessas patentes.
16. An integrated big data analytics-enabled transformation model: Application to health care	2018	Modelo conceitual de transformação habilitada para análise *big data*.
17. Simultaneously aided diagnosis model for outpatient departments via healthcare big data analytics	2018	Modelo preditivo baseado em aprendizado de máquina que auxilia médicos ambulatoriais na realização de diagnósticos.

Fonte: Autoria própria (2022)

Esta pesquisa contribui em diferentes perspectivas.

Primeiro, em considerar por meio da revisão sistemática de literatura uma breve conexão dos modelos preditivos aplicados a cuidados de saúde com um viés de TT; em sua totalidade, considera TT na perspectiva de eficácia do diagnóstico e tratamento, priorizando o viés econômico. Já as diversas pesquisas anteriores não clarificam a TT, e sim a descrevem como consequência de um processo anterior; normalmente sendo a predição.

Segundo: o modelo preditivo proposto nesta pesquisa evidencia um discurso invertido, no sentido de que a TT deve socorrer durante a apuração da predição de ressubmissão clínica. Esta necessidade se dá pelo fato de as predições produzirem descobertas significativas acerca do diagnóstico e tratamento.

Finalmente, essa investigação destaca foco e resultados distintos das pesquisas anteriores, tendo em vista a concepção analítica sobre dados e informações reais de pacientes com recebimento de diagnóstico e tratamento de

câncer e que podem ser ressubmetidos ao serviço de oncologia, devido a uma condição probabilística.

5 CONSIDERAÇÕES FINAIS

Esta pesquisa desenvolveu um modelo preditivo com funcionalidades da análise *big data* e TT, com um foco em pacientes diagnosticados com câncer e que receberam primeiro tratamento no serviço de oncologia do hospital investigado, propondo a construção da taxa de ressubmissão clínica.

Para realizar o objetivo geral, foi proposto um modelo preditivo, baseado em algoritmos de aprendizagem supervisionada, a saber, regressão logística e árvore de decisão. O modelo preditivo consiste em cinco estágios definidos: captura de dados, processamento, análise dos dados, descoberta do conhecimento e tomada de decisão e transferência de tecnologia. Um novo *dataset* foi constituído com o intuito de aplicar o modelo preditivo, considerando preditores pré-selecionados, estabelecendo os experimentos que permitiram as predições.

Em reposta aos OE 1 e OE 2 – *determinar as principais características da análise* big data *e Transferência de Tecnologia (TT) no contexto de cuidados de saúde;* e *apresentar formas de aplicação de aprendizado de máquina voltadas à análise de grandes conjuntos de dados* –, a revisão sistemática de literatura descobriu 800 documentos relacionados às especificidades das combinações das palavras-chave. Com a contribuição da *Methodi Ordinatio*, foi possível selecionar de maneira criteriosa 173 artigos, dos quais 121 foram minerados e utilizados para caracterizar a análise *big data* e TT, além de destacar todas as formas de aprendizado de máquina e suas aplicações.

Acerca do OE 3 – *construir* dataset *organizado pelos preditores estabelecidos com dados reais de pacientes submetidos a cuidados de saúde* –, a captura dos dados ocorreu no hospital selecionado, onde foram apurados 3.095 pacientes atendidos pelo serviço de oncologia entre os anos de 2010 e 2019; de tais pacientes, 41 se caracterizam como preditores preestabelecidos. Esses dados foram processados e normatizados, sendo elemento fundamental para aplicação do modelo preditivo e expansivo aos testes de diagnósticos e medidas de desempenho.

Para os OE 4 e OE 5 – *elaborar modelo preditivo com funcionalidades de análise* big data *e TT apresentando o sistema de aprendizagem em saúde, orientado a análises preditivas;* e *analisar o desempenho do modelo preditivo por meio de conjunto de dados reais de pacientes acometidos por câncer, com o foco nas medidas de desempenho* –, o modelo é desenvolvido com as funcionalidades de análise *big*

data e TT, composto por cinco estágios, contemplando algoritmos de regressão logística e árvore de decisão. Este modelo foi aplicado ao *dataset* desenvolvido, permitindo a aplicação dos testes de diagnósticos e medidas de desempenho.

Os resultados mostram que o modelo possui bom desempenho: com AUC de 0,890 para conjunto treinamento, 0,886 conjunto teste modelo II; AUC 0,886 conjunto treinamento, 0,850 para conjunto teste modelo III, para o algoritmo árvore de decisão, sensibilidade 0.897, especificidade 0.352 para conjunto de treinamento 75%, para o conjunto teste 25%, sensibilidade 0.897, especificidade 0.38. Estes valores destacam significativas correlações produzidas entre os preditores, servindo para produzir hipóteses sobre formas de adaptar a prática clínica atual e otimizar procedimentos e resultados.

Esta pesquisa possui limitações: no sentido da volatilidade dos dados dos pacientes; seleção dos preditores que colocam condições dos pacientes sobre o surgimento da doença; modelo preditivo em si, em que foi considerado sobre algoritmos de aprendizagem supervisionada, escopo único do hospital investigado. Como resposta a uma ampliação e incentivo a pesquisas futuras, aplicar este mesmo modelo em *dataset*, provido com dados de pacientes de outros hospitais, pode proporcionar refinamentos necessários a um modelo abrangente de predição em cuidados de saúde que considere TT como etapa essencial.

REFERÊNCIAS

ABE, Takumi; *et al*. Development of risk prediction models for incident frailty and their performance evaluation. **Preventive Medicine**, v. 153, p. 106768, 2021.

ABDELAZIZ, Ahmed; *et al*. A machine learning model for improving healthcare services on cloud computing environment. **Measurement,** v. 119, p. 117-128, 2018.

AHMED, Zeeshan; *et al*. Artificial intelligence with multi-functional machine learning platform development for better healthcare and precision medicine. **Database,** v. 2020, 2020.

AJAYI, Anuoluwapo; *et al*. Optimised big data analytics for health and safety hazards prediction in power infrastructure operations. **Safety Science,** v. 125, p. 104656, 2020.

AKHAVIAN, Reza; BEHZADAN, Amir H. Construction equipment activity recognition for simulation input modeling using mobile sensors and machine learning classifiers. **Advanced Engineering Informatics,** v. 29, n. 4, p. 867-877, 2015.

ALI, Munwar; *et al*. Semantic-k-NN algorithm: an enhanced version of traditional k-NN algorithm. **Expert Systems with Applications,** v. 151, p. 113374, 2020.

ALLAREDDY, Veerasathpurush; *et al*. Orthodontics in the era of big data analytics. **Orthodontics & Craniofacial Research,** v. 22, p. 8-13, 2019.

ALOTAIBI, Shoayee; *et al*. Sehaa: a big data analytics tool for healthcare symptoms and diseases detection using Twitter, Apache Spark, and machine learning**. Applied Sciences,** v. 10, n. 4, p. 1398, 2020.

AMALINA, Fairuz; *et al*. Blending big data analytics: review on challenges and a recent study. **IEEE Access,** v. 8, p. 3629-3645, 2019.

ANCARANI, Alessandro; *et al*. Technology acquisition and efficiency in Dubai hospitals. **Technological Forecasting and Social Change,** v. 113, p. 475-485, 2016.

ANGELILLO, Maria Teresa; *et al*. Attentional pattern classification for automatic dementia detection. **IEEE Access,** v. 7, p. 57706-57716, 2019.

ANOWAR, Farzana; SADAOUI, Samira; SELIM, Bassant. Conceptual and empirical comparison of dimensionality reduction algorithms (PCA, KPCA, LDA, MDS, SVD, LLE, ISOMAP, LE, ICA, t-SNE). **Computer Science Review,** v. 40, p. 100378, 2021.

APPOLINÁRIO, Fabio. **Dicionário de metodologia científica:** um guia para a produção do conhecimento científico. São Paulo: Atlas, 2009.

BABAR, Muhammad; *et al*. Energy-harvesting based on internet of things and big data analytics for smart health monitoring. **Sustainable Computing: Informatics and Systems,** v. 20, p. 155-164, 2018.

BATTISTONI, Giuseppe; *et al.* Cost-benefit analysis of applied research infrastructure: evidence from health care. **Technological Forecasting and Social Change,** v. 112, p. 79-91, 2016.

BEN-ASSULI, Ofir; *et al*. Bringing big data analytics closer to practice: a methodological explanation and demonstration of classification algorithms. **Health Policy and Technology,** v. 8, n. 1, p. 7-13, 2019.

BERGER, Jeffrey S. et al. Evaluation of machine learning methodology for the prediction of healthcare resource utilization and healthcare costs in patients with critical limb ischemia-is preventive and personalized approach on the horizon?. **EPMA Journal**, v. 11, n. 1, p. 53-64, 2020.

BIRKEN, Sarah A.; *et al*. Elaborating on theory with middle managers' experience implementing healthcare innovations in practice. **Implementation Science,** v. 11, n. 1, p. 1-5, 2015.

BZDOK, Danilo; IOANNIDIS, John P. A. Exploration, inference, and prediction in neuroscience and biomedicine. **Trends in Neurosciences,** v. 42, n. 4, p. 251-262, 2019.

CHAN, Siu L.; LU, Yanglong; WANG, Yan. Data-driven cost estimation for additive manufacturing in cybermanufacturing. **Journal of Manufacturing Systems,** v. 46, p. 115-126, 2018.

CHAUHAN, Ritu; KAUR, Harleen; CHANG, Victor. An optimized integrated framework of big data analytics managing security and privacy in healthcare data. **Wireless Personal Communications,** p. 1-22, 2020.

CHEN, Min. *et al*. Disease prediction by machine learning over big data from healthcare communities. **IEEE Access,** v. 5, p. 8869-8879, 2017.

CHOI, Jaehyun; *et al*. A predictive model of technology transfer using patent analysis. **Sustainability,** v. 7, n. 12, p. 16175-16195, 2015.

CHRIMES, Dillon; ZAMANI, Hamid. Using distributed data over HBase in big data analytics platform for clinical services. **Computational and Mathematical Methods in Medicine,** v. 2017, 2017.

CRESWELL, John W. **Projeto de pesquisa:** métodos qualitativo, quantitativo e misto. Porto Alegre: Artmed, 2010.

DAVIES, Gareth Huw; RODERICK, Sian; HUXTABLE, Louisa Thomas. Social commerce open innovation in healthcare management: an exploration from a novel technology transfer approach. **Journal of Strategic Marketing,** v. 27, n. 4, p. 356-367, 2019.

DE MAURO, Andrea; *et al.* Human resources for big data professions: a systematic classification of job roles and required skill sets. **Information Processing & Management,** v. 54, n. 5, p. 807-817, 2018.

DUBEY, Rameshwar; *et al*. Big data analytics and artificial intelligence pathway to operational performance under the effects of entrepreneurial orientation and environmental dynamism: a study of manufacturing organisations. **International Journal of Production Economics,** v. 226, p. 107599, 2020.

FAN, Cheng et al. A novel methodology to explain and evaluate data-driven building energy performance models based on interpretable machine learning. **Applied Energy**, v. 235, p. 1551-1560, 2019.

FERNÁNDEZ, Antonio M.; *et al*. Automated deployment of a spark cluster with machine learning algorithm integration. Big data **Research,** v. 19, p. 100135, 2020.

GAO, Kaifeng et al. Julia language in machine learning: Algorithms, applications, and open issues. **Computer Science Review**, v. 37, p. 100254, 2020.

GALETSI, Panagiota; KATSALIAKI, Korina; KUMAR, Sameer. Big data analytics in health sector: theoretical framework, techniques and prospects. **International Journal of Information Management,** v. 50, p. 206-216, 2020.

GARCÍA, Antonio J. et al. Automatic alarm prioritization by data mining for fault management in cellular networks. **Expert Systems with Applications**, v. 158, p. 113526, 2020.

GARTNER, Daniel; PADMAN, Rema. Machine learning for healthcare behavioural OR: Addressing waiting time perceptions in emergency care. **Journal of the Operational Research Society**, v. 71, n. 7, p. 1087-1101, 2020.

GASIMOVA, Rena T.; ABBASLI, Rahim N. Advancement of the search process for digital heritage by utilizing artificial intelligence algorithms. **Expert Systems with Applications,** v. 158, p. 113559, 2020.

GHANI, Norjihan Abdul; *et al.* Social media big data analytics: a survey. **Computers in Human Behavior,** v. 101, p. 417-428, 2019.

GIL, Antonio Carlos. **Métodos e técnicas de pesquisa social.** 5. ed. São Paulo: Atlas, 1999.

GOMES, Myller Augusto Santos; *et al*. Government initiative in Brazilian public health: a technology transfer analysis. **International Journal of Environmental Research and Public Health**, v. 16, n. 17, p. 3012, 2019.

GUO, Chonghui; CHEN, Jingfeng. Big data analytics in healthcare: data-driven methods for typical treatment pattern mining. **Journal of Systems Science and Systems Engineering,** v. 28, n. 6, p. 694-714, 2019.

HADI, Mohammed S; *et al*. Patient-centric HetNets powered by machine learning and big data analytics for 6G networks. **IEEE Access**, v. 8, p. 85639-85655, 2020.

HAQ, Amin Ul; *et al*. Intelligent machine learning approach for effective recognition of diabetes in e-healthcare using clinical data. **Sensors,** v. 20, n. 9, p. 2649, 2020.

HARERIMANA, Gaspard; *et al*. Health big data analytics: a technology survey. **IEEE Access,** v. 6, p. 65661-65678, 2018.

HATTON, Christopher M; *et al*. Predicting persistent depressive symptoms in older adults: a machine learning approach to personalised mental healthcare. **Journal of affective disorders**, v. 246, p. 857-860, 2019.

HIDALGO, Antonio; *et al*. The digital divide in light of sustainable development: an approach through advanced machine learning techniques. **Technological Forecasting and Social Change,** v. 150, p. 119754, 2020.

HU, Ying et al. Simultaneously aided diagnosis model for outpatient departments via healthcare big data analytics. **Multimedia Tools and Applications**, v. 77, n. 3, p. 3729-3743, 2018.

HUANG, Ying; *et al*. A technology delivery system for characterizing the supply side of technology emergence: illustrated for big data & analytics. **Technological Forecsting and Social Change,** v. 130, p. 165-176, 2018.

IP, Ryan Ho Leung; *et al*. Big data and machine learning for crop protection. **Computers and Electronics in Agriculture,** v. 151, p. 376-383, 2018.

ISTEPANIAN, Robert S. H.; AL-ANZI, Turki. m-Health 2.0: new perspectives on mobile health, machine learning and big data analytics. **Methods,** v. 151, p. 34-40, 2018.

JADHAV, Suresh; GAUTAM, Manish; GAIROLA, Sunil. Role of vaccine manufacturers in developing countries towards global healthcare by providing quality

vaccines at affordable prices. **Clinical Microbiology and Infection,** v. 20, p. 37-44, 2014.

JENA, Rabindra. An empirical case study on Indian consumers' sentiment towards electric vehicles: a big data analytics approach. **Industrial Marketing Management,** v. 90, p. 605-616, oct. 2020.

JOHNSTON, Stephen S.; *et al*. Using machine learning applied to real-world healthcare data for predictive analytics: an applied example in bariatric surgery. **Value in Health,** v. 22, n. 5, p. 580-586, 2019.

JONES, Luke D.; *et al*. Artificial intelligence, machine learning and the evolution of healthcare: a bright future or cause for concern? **Bone & Joint Research,** v. 7, n. 3, p. 223-225, 2018.

JUNIOR, Warley; *et al*. A context-sensitive offloading system using machine-learning classification algorithms for mobile cloud environment. **Future Generation Computer Systems**, v. 90, p. 503-520, 2019.

KAPOOR, Rahul; LEE, Joon Mahn. Coordinating and competing in ecosystems: how organizational forms shape new technology investments. **Strategic Management Journal,** v. 34, n. 3, p. 274-296, 2013.

KASHI, Shir *et al*. A machine-learning model for automatic detection of movement compensations in stroke patients. **IEEE Transactions on Emerging Topics in Computing,** 23 apr. 2020.

KEMPA-LIEHR, Andreas W. et al. Healthcare pathway discovery and probabilistic machine learning. **International journal of medical informatics**, v. 137, p. 104087, 2020.

KHAN, Muhammad Ashfaq; *et al*. A two-stage big data analytics framework with real world applications using spark machine learning and long short-term memory network. **Symmetry,** v. 10, n. 10, p. 485, 2018.

KIBRIA, Mirza Golam *et al*. Big data analytics, machine learning, and artificial intelligence in next-generation wireless networks. **IEEE Access,** v. 6, p. 32328-32338, 2018.

KOSE, Ilker; GOKTURK, Mehmet; KILIC, Kemal. An interactive machine-learning-based electronic fraud and abuse detection system in healthcare insurance. **Applied Soft Computing,** v. 36, p. 283-299, 2015.

KOZJEK, Dominik; *et al.* Advancing manufacturing systems with big-data analytics: a conceptual framework. **International Journal of Computer Integrated Manufacturing,** v. 33, n. 2, p. 169-188, 2020.

KWAKERNAAK, Sascha; *et al*. Using machine learning to predict mental healthcare consumption in non-affective psychosis. **Schizophrenia Research,** v. 218, p. 166-172, 2020.

KWON, Jung-Hyok; LEE, Hwi-Ho; KIM, Eui-Jik. Big data analytics-based predictive modeling for stress management using healthcare system. **Advanced Science Letters**, v. 23, n. 3, p. 1585-1588, 2017.

LAKATOS, Eva Maria; MARCONI, Marina de Andrade. **Fundamentos de metodologia científica.** 4. ed. São Paulo: Atlas, 2001.

LANCASTER, Megan Cummins; *et al*. Phenotypic clustering of left ventricular diastolic function parameters: patterns and prognostic relevance. **JACC: Cardiovascular Imaging,** v. 12, n. 7, p. 1149-1161, 2019.

LANDIS, J. Richard; KOCH, Gary G. The measurement of observer agreement for categorical data. **biometrics**, p. 159-174, 1977.

LEE, Jay H.; SHIN, Joohyun; REALFF, Matthew J. Machine learning: overview of the recent progresses and implications for the process systems engineering field. **Computers & Chemical Engineering,** v. 114, p. 111-121, 2018.

LEITE, André Ferreira; *et al*. Radiomics and machine learning in oral healthcare. **Proteomics - Clinical Applications,** v. 14, n. 3, p. 1900040, 2020.

LI, Jian Ping; *et al*. Heart disease identification method using machine learning classification in e-healthcare. **IEEE Access,** v. 8, p. 107562-107582, 2020.

LI, Yi; *et al*. Data-driven health estimation and lifetime prediction of lithium-ion batteries: a review. **Renewable and Sustainable Energy Reviews,** v. 113, p. 109254, 2019.

LIN, Hui-Heng; OUYANG, Defang; HU, Yuanjia. Intelligent classifier: a tool to impel drug technology transfer from academia to industry. **Journal of Pharmaceutical Innovation**, v. 14, n. 1, p. 28-34, 2019.

LIMA, Marcio Salles Melo; DELEN, Dursun. Predicting and explaining corruption across countries: a machine learning approach. **Government Information Quarterly,** v. 37, n. 1, p. 101407, 2020.

LIU, Mengchen; *et al*. Analyzing the training processes of deep generative models. **IEEE Transactions on Visualization and Computer Graphics,** v. 24, n. 1, p. 77-87, 2017.

LIU, Shixia *et al*. Towards better analysis of machine learning models: A visual analytics perspective. **Visual Informatics**, v. 1, n. 1, p. 48-56, 2017.

MA, Chuang; ZHANG, Hao Helen; WANG, Xiangfeng. Machine learning for big data analytics in plants. **Trends in Plant Science,** v. 19, n. 12, p. 798-808, 2014.

MAZUMDAR, Madhu; *et al*. Comparison of statistical and machine learning models for healthcare cost data: a simulation study motivated by oncology care model (OCM) data. **BMC Health Services Research,** v. 20, p. 1-12, 2020.

MCNUTT, Todd R.; *et al*. Using big data analytics to advance precision radiation oncology. **International Journal of Radiation Oncology Biology Physics,** v. 101, n. 2, p. 285-291, 2018.

MCWILLIAMS, Christopher J; *et al*. Towards a decision support tool for intensive care discharge: machine learning algorithm development using electronic healthcare data from MIMIC-III and Bristol, UK. **BMJ open**, v. 9, n. 3, p. e025925, 2019.

MEHTA, Nishita; PANDIT, Anil. Concurrence of big data analytics and healthcare: A systematic review. **International Journal of Medical Informatics,** v. 114, p. 57-65, 2018.

MEHTA, Nishita; PANDIT, Anil; SHUKLA, Sharvari. Transforming healthcare with big data analytics and artificial intelligence: a systematic mapping study. **Journal of Biomedical Informatics,** v. 100, p. 103311, 2019.

MILLER, Fiona A.; FRENCH, Martin. Organizing the entrepreneurial hospital: hybridizing the logics of healthcare and innovation. **Research Policy,** v. 45, n. 8, p. 1534-1544, 2016.

MOHAMED, Azlinah; *et al*. The state of the art and taxonomy of big data analytics: view from new big data framework. **Artificial Intelligence Review,** v. 53, n. 2, p. 989-1037, 2020.

MOREIRA, Mario Wedney de Lima; *et al*. Postpartum depression prediction through pregnancy data analysis for emotion-aware smart systems. **Information Fusion,** v. 47, p. 23-31, 2019.

MOYO, Sangiwe; *et al*. Application of machine learning models in predicting length of stay among healthcare workers in underserved communities in South Africa. **Human Resources for Health,** v. 16, n. 1, p. 1-9, 2018.

NAIR, Lekha R.; SHETTY, Sujala D.; SHETTY, Siddhanth D. Applying spark based machine learning model on streaming big data for health status prediction. **Computers & Electrical Engineering,** v. 65, p. 393-399, 2018.

NICOL, Dianne; *et al*. Understanding public reactions to commercialization of biobanks and use of biobank resources. **Social Science & Medicine,** v. 162, p. 79-87, 2016.

NILSEN, Etty R.; *et al*. Exploring resistance to implementation of welfare technology in municipal healthcare services: a longitudinal case study. **BMC Health Services Research,** v. 16, n. 1, p. 1-14, 2016.

PADARIAN, José; MINASNY, Budiman; MCBRATNEY, Alex B. Using deep learning to predict soil properties from regional spectral data. **Geoderma Regional,** v. 16, p. e00198, 2019.

PAGANI, Regina Negri; KOVALESKI, João Luiz; RESENDE, Luis Mauricio Martins de. *Methodi Ordinatio*: a proposed methodology to select and rank relevant scientific papers encompassing the impact factor, number of citation, and year of publication. **Scientometrics,** v. 105, n. 3, p. 2109-2135, 2015.

PAN, Jinxin; *et al*. Exploring behavioural intentions toward smart healthcare services among medical practitioners: a technology transfer perspective. **International Journal of Production Research,** v. 57, n. 18, p. 5801-5820, 2019.

PANAYIDES, Andreas S.; *et al*. Radiogenomics for precision medicine with a big data analytics perspective. **IEEE Journal of Biomedical and Health Informatics,** v. 23, n. 5, p. 2063-2079, 2018.

PASSOS, Ives C.; *et al*. Machine learning and big data analytics in bipolar disorder: a position paper from the International Society for Bipolar Disorders Big data Task Force. **Bipolar Disorders,** v. 21, n. 7, p. 582-594, 2019.

PRIHODOVA, Lucia; GUERIN, Suzanne; KERNOHAN, W. George. Knowledge transfer and exchange frameworks in health and their applicability to palliative care: scoping review protocol. **Journal of Advanced Nursing,** v. 71, n. 7, p. 1717-1725, 2015.

QAISAR, Saeed Mian; SUBASI, Abdulhamit. Effective epileptic seizure detection based on the event-driven processing and machine learning for mobile healthcare. **Journal of Ambient Intelligence and Humanized Computing**, p. 1-13, 2020.

RAZZAK, Muhammad Imran; IMRAN, Muhammad; XU, Guandong. Big data analytics for preventive medicine. **Neural Computing and Applications,** v. 32, n. 9, p. 4417-4451, 2020.

REPS, Jenna; *et al*. Design and implementation of a standardized framework to generate and evaluate patient-level prediction models using observational healthcare data. **Journal of the American Medical Informatics Association**, v. 25, n. 8, p. 969-975, 2018.

RISTEVSKI, Blagoj; CHEN, Ming. Big data analytics in medicine and healthcare. **Journal of Integrative Bioinformatics,** v. 15, n. 3, 2018.

RIZWAN, Ali; *et al*. A review on the role of nano-communication in future healthcare systems: a big data analytics perspective. **IEEE Access,** v. 6, p. 41903-41920, 2018.

ROTH, Jan A.; *et al*. Introduction to machine learning in digital healthcare epidemiology. **Infection Control & Hospital Epidemiology,** v. 39, n. 12, p. 1457-1462, 2018.

SAGGI, Mandeep Kaur; JAIN, Sushma. A survey towards an integration of big data analytics to big insights for value-creation. **Information Processing & Management,** v. 54, n. 5, p. 758-790, 2018.

SAHEB, Tahereh; IZADI, Leila. Paradigm of IoT big data analytics in the healthcare industry: a review of scientific literature and mapping of research trends. **Telematics and Informatics,** v. 41, p. 70-85, 2019.

SAVIN, Ivan; *et al.* Healthcare-associated ventriculitis and meningitis in a neuro-ICU: incidence and risk factors selected by machine learning approach. **Journal of Critical Care,** v. 45, p. 95-104, 2018.

SHAFQAT, Sarah; *et al*. Big data analytics enhanced healthcare systems: a review. **The Journal of Supercomputing,** v. 76, n. 3, p. 1754-1799, 2020.

SHARMA, Rohit; *et al*. A systematic literature review on machine learning applications for sustainable agriculture supply chain performance. **Computers & Operations Research,** v. 119, p. 104926, 2020.

SIVAPARTHIPAN, C. B.; *et al*. Innovative and efficient method of robotics for helping the Parkinson's disease patient using IoT in big data analytics. **Transactions on Emerging Telecommunications Technologies,** v. 31, n. 12, p. e3838, 2020.

SOURI, Alireza; *et al.* A new machine learning-based healthcare monitoring model for student's condition diagnosis in Internet of Things environment. **Soft Computing,** v. 24, p. 17111-17121, 2020.

SOUSA, Maria José et al. Decision-making based on big data analytics for people management in healthcare organizations. **Journal of medical systems**, v. 43, n. 9, p. 1-10, 2019.

SRIVASTAVA, Saurabh Kumar; SINGH, Sandeep Kumar; SURI, Jasjit S. Effect of incremental feature enrichment on healthcare text classification system: A machine learning paradigm. **Computer methods and programs in biomedicine**, v. 172, p. 35-51, 2019.

STRANG, Kenneth David; SUN, Zhaohao. Hidden big data analytics issues in the healthcare industry. **Health Informatics Journal,** v. 26, n. 2, p. 981-998, 2020.

SUBUDHI, Badri Narayan; ROUT, Deepak Kumar; GHOSH, Ashish. Big data analytics for video surveillance. **Multimedia Tools and Applications,** v. 78, n. 18, p. 26129-26162, 2019.

SUJITHA, R.; SEENIVASAGAM, V. Classification of lung cancer stages with machine learning over big data healthcare framework. **Journal of Ambient Intelligence and Humanized Computing**, p. 1-11, 2020.

SUN, Shaolong *et al*. Forecasting tourist arrivals with machine learning and internet search index. Tourism Management, v. 70, p. 1-10, 2019.

SUTHAHARAN, Shan. Big data analytics: machine learning and bayesian learning perspectives: what is done? what is not? **Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery,** v. 9, n. 1, p. e1283, 2019.

SYED, Liyakathunisa et al. Smart healthcare framework for ambient assisted living using IoMT and big data analytics techniques. **Future Generation Computer Systems**, v. 101, p. 136-151, 2019.

TABESH, Pooya; MOUSAVIDIN, Elham; HASANI, Sona. Implementing big data strategies: a managerial perspective. **Business Horizons,** v. 62, n. 3, p. 347-358, 2019.

THARWAT, Alaa; *et al*. Linear discriminant analysis: A detailed tutorial. **AI communications**, v. 30, n. 2, p. 169-190, 2017.

VENKATESAN, Rajagopal; SRINIVASAN, Balakrishnan; RAJENDIRAN, Periyasamy. tiger hash based AdaBoost machine learning classifier for secured multicasting in mobile healthcare system. **Cluster Computing,** v. 22, n. 3, p. 7039-7053, 2019.

WANG, Yichuan; HAJLI, Nick. Exploring the path to big data analytics success in healthcare. **Journal of Business Research,** v. 70, p. 287-299, 2017.

WANG, Yichuan; KUNG, Lee Ann; BYRD, Terry Anthony. Big data analytics: understanding its capabilities and potential benefits for healthcare organizations. **Technological Forecasting and Social Change,** v. 126, p. 3-13, 2018.

WANG, Yichuan; *et al*. An integrated big data analytics-enabled transformation model: application to healthcare. **Information & Management,** v. 55, n. 1, p. 64-79, 2018.

WARING, Jonathan; LINDVALL, Charlotta; UMETON, Renato. Automated machine learning: review of the state-of-the-art and opportunities for healthcare. **Artificial Intelligence in Medicine,** p. 101822, 2020.

WIENS, Jenna; SHENOY, Erica S. Machine learning for healthcare: on the verge of a major shift in healthcare epidemiology. **Clinical Infectious Diseases,** v. 66, n. 1, p. 149-153, 2018.

YACCHIREMA, Diana C.; *et al*. A smart system for sleep monitoring by integrating IoT with big data analytics. **IEEE Access,** v. 6, p. 35988-36001, 2018.

YU, Yiding; WANG, Taotao; LIEW, Soung Chang. Deep-reinforcement learning multiple access for heterogeneous wireless networks. **IEEE Journal on Selected Areas in Communications,** v. 37, n. 6, p. 1277-1290, 2019.

YUVARAJ, N.; SRIPREETHAA, K. R. Diabetes prediction in healthcare systems using machine learning algorithms on Hadoop cluster. **Cluster Computing,** v. 22, n. 1, p. 1-9, 2019.

ZHANG, Caifeng; *et al*. Optimizing the electronic health records through big data analytics: a knowledge-based view. **IEEE Access,** v. 7, p. 136223-136231, 2019.

ZHANG, Yi; *et al*. Discovering and forecasting interactions in big data research: A learning-enhanced bibliometric study. **Technological Forecasting and Social Change,** v. 146, p. 795-807, 2019.

ZUBAR, A. H.; BALAMURUGAN, R. Green computing process and its optimization using machine learning algorithm in healthcare sector. **Mobile Networks and Applications,** v. 25, p. 1307-1318, 2020.

APÊNDICE A - Tabela de aplicação do *Methodi Ordinatio*

Id	Título	Ano	Citações	FI	*In Ordinatio*
1	Big data analytics: Understanding its capabilities and potential benefits for healthcare organizations	2018	530	3,815	610,003815
2	Disease Prediction by Machine Learning Over Big Data From Healthcare Communities	2017	334	3,745	404,00
3	An integrated big data analytics-enabled transformation model: Application to health care	2018	155	4,12	235,00412
4	Exploring the path to big data analytics success in healthcare	2017	161	4,028	231,004028
5	Linear discriminant analysis: A detailed tutorial	2017	137	0,765	207,000765
6	-Omic and Electronic Health Record Big Data Analytics for Precision Medicine	2017	128	4,491	198,004491
7	Concurrence of big data analytics and healthcare: A systematic review	2018	111	2,731	191,002731
8	Big Data: Tutorial and guidelines on information and process fusion for analytics algorithms with MapReduce	2018	96	10,716	176,010716
9	Big Data Analytics, Machine Learning, and Artificial Intelligence in Next-Generation Wireless Networks	2018	96	4,098	176,004098
10	Technology: The Future of Agriculture	2017	104	43,070	174,04307
11	Deep-Reinforcement Learning Multiple Access for Heterogeneous Wireless Networks	2019	84	9,302	174,009302
12	Computer vision and deep learning techniques for pedestrian detection and tracking: A survey	2018	87	4,072	167,004072
13	Machine learning: Overview of the recent progresses and implications for the process systems engineering field	2018	83	4	163,004
14	Big data analytics for disaster response and recovery through sentiment analysis	2018	82	5,063	162,005063
15	A survey towards an integration of big data analytics to big insights for value-creation	2018	75	4.787	159,787
16	A machine learning model for improving healthcare services on cloud computing environment	2018	75	3,364	155,00
17	Analyzing the Training Processes of Deep Generative Models	2018	73	3,78	153,00378
18	Big Data Analytics for Genomic Medicine	2017	81	4,183	151,004183
19	Machine Learning for Healthcare: On the Verge of a Major Shift in Healthcare Epidemiology	2018	67	8,313	147,01
20	Human resources for Big Data professions: A systematic classification of job roles and required skill sets	2018	66	3,889	146,003889

21	Forecasting tourist arrivals with machine learning and internet search index	2019	53	6,012	143,006012
22	Social media big data analytics: A survey	2019	44	4,306	134,004306
23	Data-driven decision making under uncertainty integrating robust optimization with principal component analysis and kernel smoothing methods	2018	54	3,334	134,003
24	Smart urban planning using Big Data analytics to contend with the interoperability in Internet of Things	2017	54	5,768	124,005768
25	Patient-Centric Cellular Networks Optimization Using Big Data Analytics	2019	34	4,098	124,004098
26	ClowdFlows: Online workflows for distributed big data mining	2017	53	5,768	123,005768
27	Exploration, Inference, and Prediction in Neuroscience and Biomedicine	2019	32	12,314	122,012314
28	Data-driven health estimation and lifetime prediction of lithium-ion batteries: A review	2019	31	10,556	121,010556
29	Data-driven cost estimation for additive manufacturing in cybermanufacturing	2018	40	3,642	120,003642
30	Big data analytics enhanced healthcare systems: a review	2020	18	2,157	118,00
31	A big data driven sustainable manufacturing framework for condition-based maintenance prediction	2018	36	2,502	116,002502
32	Using deep learning to predict soil properties from regional spectral data	2019	26	1,5	116,0015
33	Smart health monitoring and management system: Toward autonomous wearable sensing for Internet of Things using big data analytics	2019	25	5,768	115,005768
34	Big data analytics for preventive medicine	2020	15	4,664	115,004664
35	Big data analytics in health sector: Theoretical framework, techniques and prospects	2020	13	5,063	113,005063
36	A Smart System for Sleep Monitoring by Integrating IoT With Big Data Analytics	2018	33	4,098	113,004098
37	Molecular pathway activation - New type of biomarkers for tumor morphology and personalized selection of target drugs	2018	32	9,658	112,009658
38	m-Health 2.0: New perspectives on mobile health, machine learning and big data analytics	2018	32	3,782	112,003782
39	Implementing big data strategies: A managerial perspective	2019	22	2,828	112,002828
40	The state of the art and taxonomy of big data analytics: view from new big data framework	2020	10	5,095	110,005095
41	Artificial intelligence, machine learning and the evolution of healthcare: A bright future or cause for concern?	2018	29	3,532	109,00

42	Big data analytics for personalized medicine	2019	18	8,083	108,008083
43	Preference learning for eco-friendly hotels recommendation: A multi-criteria collaborative filtering approach	2019	18	6,395	108,006395
44	Big Data Analytics of Identifying Geochemical Anomalies Supported by Machine Learning Methods	2018	28	2	108,002
45	Times-series data augmentation and deep learning for construction equipment activity recognition	2019	16	3,772	106,003772
46	Paradigm of IoT big data analytics in the healthcare industry: A review of scientific literature and mapping of research trends	2019	16	3,714	106,003714
47	Diabetes prediction in healthcare systems using machine learning algorithms on Hadoop cluster	2019	16	3,458	106,00
48	Big data and machine learning for crop protection	2018	26	3,171	106,003171
49	Simultaneously aided diagnosis model for outpatient departments via healthcare big data analytics	2018	26	2,101	106,002101
50	The ICO and artificial intelligence: The role of fairness in the GDPR framework	2018	26	1,552	106,001552
51	Optimized Scoring Systems: Toward Trust in Machine Learning for Healthcare and Criminal Justice	2018	26	0,775	106,00
52	A hybrid IT framework for identifying high-quality physicians using big data analytics	2019	15	5,063	105,005063
53	Big data analytics and artificial intelligence pathway to operational performance under the effects of entrepreneurial orientation and environmental dynamism: A study of manufacturing organisations	2020	5	4,998	105,004998
54	Automated machine learning: Review of the state-of-the-art and opportunities for healthcare	2020	5	3,574	105,00
55	Big Data Systems Meet Machine Learning Challenges: Towards Big Data Science as a Service	2018	25	2,952	105,002952
56	Leveraging Big Data Analytics to Improve Quality of Care in Healthcare Organizations: A Configurational Perspective	2019	15	2,75	105,00275
57	Towards a real-time processing framework based on improved distributed recurrent neural network variants with fastText for social big data analytics	2020	4	3,889	104,003889
58	The digital divide in light of sustainable development: An approach through advanced machine learning techniques	2020	4	3,815	104,003815

59	Artificial intelligence with multi-functional machine learning platform development for better healthcare and precision medicine	2020	4	2,593	104,00
60	Radiomics and Machine Learning in Oral Healthcare	2020	4	2,324	104,00
61	Predicting and explaining corruption across countries: A machine learning approach	2020	3	4,311	103,004311
62	Predicting persistent depressive symptoms in older adults: A machine learning approach to personalised mental healthcare	2019	13	4,084	103,00
63	Iot based laundry services: an application of big data analytics, intelligent logistics management, and machine learning techniques	2020	3	3,199	103,003199
64	Earthquake prediction in California using regression algorithms and cloud-based big data infrastructure	2018	23	2,721	103,002721
65	Machine learning and data mining advance predictive big data analysis in precision animal agriculture	2018	23	1,697	103,001697
66	A novel methodology to explain and evaluate data-driven building energy performance models based on interpretable machine learning	2019	12	8,426	102,01
67	Intelligent machine learning approach for effective recognition of diabetes in e-healthcare using clinical data	2020	2	3,275	102,003275
68	Urban Planning and Smart City Decision Management Empowered by Real-Time Data Processing Using Big Data Analytics	2018	22	3,031	102,003031
69	A systematic literature review on machine learning applications for sustainable agriculture supply chain performance	2020	2	3,002	102,00
70	Teleconsultations between patients and healthcare professionals in primary care in catalonia: The evaluation of text classification algorithms using supervised machine learning	2020	2	2,849	102,003
71	P2P-based open health cloud for medicine management	2020	2	2,793	102,00
72	Sehaa: A Big Data Analytics Tool for Healthcare Symptoms and Diseases Detection Using Twitter, Apache Spark, and Machine Learning	2020	2	2,474	102,002474
73	An empirical case study on Indian consumers' sentiment towards electric vehicles: A big data analytics approach	2020	1	4,797	101,004797

74	Using machine learning to predict mental healthcare consumption in non-affective psychosis	2020	1	4,569	101,004569
75	Providing Healthcare-as-a-Service Using Fuzzy Rule Based Big Data Analytics in Cloud Computing	2018	21	4,217	101,004217
76	Patient-Centric HetNets Powered by Machine Learning and Big Data Analytics for 6G Networks	2020	1	4,098	101,004098
77	Fast and scalable distributed deep convolutional autoencoder for fMRI big data analytics	2019	11	4,072	101,004072
78	The role of location and social strength for friendship prediction in location-based social networks	2018	21	3,889	101,00
79	Measuring the effects of confounders in medical supervised classification problems: the Confounding Index (CI)	2020	1	3,574	101,004
80	A new machine learning-based healthcare monitoring model for student's condition diagnosis in Internet of Things environment	2020	1	3,050	101,00
81	A technology-aided multi-modal training approach to assist abdominal palpation training and its assessment in medical education	2020	1	2,006	101,002006
82	Comparison of statistical and machine learning models for healthcare cost data: a simulation study motivated by Oncology Care Model (OCM) data	2020	1	1,987	101,00
83	Phenotypic Clustering of Left Ventricular Diastolic Function Parameters: Patterns and Prognostic Relevance	2018	20	10,975	100,01
84	Postpartum depression prediction through pregnancy data analysis for emotion-aware smart systems	2019	10	10,716	100,010716
85	Machine learning as a tool to design glasses with controlled dissolution for healthcare applications	2020	0	6,638	100,006638
86	A hybrid machine learning framework for analyzing human decision-making through learning preferences	2020	0	5,341	100,005341
87	Radiogenomics for Precision Medicine with a Big Data Analytics Perspective	2019	10	5,223	100,01
88	A machine-learning model for automatic detection of movement compensations in stroke patients	2020	0	4,989	100,004989
89	Evaluation of machine learning methodology for the prediction of healthcare resource utilization and healthcare costs in patients with critical limb ischemia-is preventive and personalized approach on the horizon?	2020	0	4,901	100,00

90	Mapping hurricane damage: A comparative analysis of satellite monitoring methods	2020	0	4,846	100,004846
91	Effective epileptic seizure detection based on the event-driven processing and machine learning for mobile healthcare	2020	0	4,594	100,004594
92	Advancement of the search process for Digital Heritage by utilizing artificial intelligence algorithms	2020	0	4,292	100,004292
93	Automatic alarm prioritization by data mining for fault management in cellular networks	2020	0	4,292	100,004292
94	Scalable auto-encoders for gravitational waves detection from time series data	2020	0	4,292	100,004292
95	Learning alternative ways of performing a task	2020	0	4,292	100,004292
96	Heart Disease Identification Method Using Machine Learning Classification in E-Healthcare	2020	0	3,745	100,003745
97	Optimised Big Data analytics for health and safety hazards prediction in power infrastructure operations	2020	0	3,619	100,00
98	Automated Deployment of a Spark Cluster with Machine Learning Algorithm Integration	2020	0	2,952	100,003
99	Social media based surveillance systems for healthcare using machine learning: A systematic review	2020	0	2,95	100,00
100	Automated myocardial infarction identification based on interbeat variability analysis of the photoplethysmographic data	2020	0	2,943	100,002943
101	Healthcare pathway discovery and probabilistic machine learning	2020	0	2,731	100,00
102	Green Computing Process and its Optimization Using Machine Learning Algorithm in Healthcare Sector	2020	0	2,602	100,00
103	Moderating Effects of Gender and Resistance to Change on the Adoption of Big Data Analytics in Healthcare	2020	0	2,591	100,002591
104	Self-Service Data Science in Healthcare with Automated Machine Learning	2020	0	2,474	100,00
105	Big data analytics and processing platform in Czech Republic healthcare	2020	0	2,217	100,002217
106	Assessment of utilization efficiency using machine learning techniques: A study of heterogeneity in preoperative healthcare utilization among super-utilizers	2020	0	2,201	100,00
107	Advancing manufacturing systems with big-data analytics: A conceptual framework	2020	0	2,09	100,00209

108	Personalised healthcare model for monitoring and prediction of airpollution: machine learning approach	2020	0	2,039	100,002039
109	A review of machine learning for big data analytics: bibliometric approach	2020	0	1,739	100,00
110	Using Applied Machine Learning to Predict Healthcare Utilization Based on Socioeconomic Determinants of Care	2020	0	1,4	100,0014
111	An Optimized Integrated Framework of Big Data Analytics Managing Security and Privacy in Healthcare Data	2020	0	0,929	100,000929
112	A distributed evolutionary multivariate discretizer for Big Data processing on Apache Spark	2018	19	6,33	99,00633
113	Energy-efficient hadoop for big data analytics and computing: A systematic review and research insights	2018	19	5,768	99,006
114	A context-sensitive offloading system using machine-learning classification algorithms for mobile cloud environment	2019	9	5,768	99,005768
115	A Review on the Role of Nano-Communication in Future Healthcare Systems: A Big Data Analytics Perspective	2018	19	4,098	99,004098
116	Exploring behavioural intentions toward smart healthcare services among medical practitioners: a technology transfer perspective	2019	8	4,577	98,004577
117	A big data driven distributed density based hesitant fuzzy clustering using Apache spark with application to gene expression microarray	2019	8	3,526	98,003526
118	Smart healthcare framework for ambient assisted living using IoMT and big data analytics techniques	2019	7	5,768	97,005768
119	Machine learning and big data analytics in bipolar disorder: A position paper from the International Society for Bipolar Disorders Big Data Task Force	2019	7	4,936	97,005
120	Discovering and forecasting interactions in big data research: A learning-enhanced bibliometric study	2019	7	3,815	97,003815
121	Effect of incremental feature enrichment on healthcare text classification system: A machine learning paradigm	2019	7	3,424	97,00
122	Values, challenges and future directions of big data analytics in healthcare: A systematic review	2019	7	3,087	97,003087
123	Towards a decision support tool for intensive care discharge: machine learning algorithm development using electronic healthcare data from MIMIC-III and Bristol, UK	2019	7	2,496	97,00

124	Energy-harvesting based on internet of things and big data analytics for smart health monitoring	2018	17	1,8	97,0018
125	Blending Big Data Analytics: Review on Challenges and a Recent Study	2019	6	4,098	96,004098
126	Health Big Data Analytics: A Technology Survey	2018	16	4,098	96,004098
127	Attentional Pattern Classification for Automatic Dementia Detection	2019	6	4,098	96,004098
128	Designing medical technology for resilience: integrating health economics and human factors approaches	2018	15	2,2	95,002
129	A review of the literature on big data analytics in healthcare	2019	5	1,754	95,001754
130	Using Machine Learning Applied to Real-World Healthcare Data for Predictive Analytics: An Applied Example in Bariatric Surgery	2019	4	5,037	94,01
131	Agent-Based Simulation of Smart Beds With Internet-of-Things for Exploring Big Data Analytics	2018	14	4,098	94,004098
132	A 6-month analysis of factors impacting web browsing quality for QoE prediction	2019	4	3,03	94,00303
133	Machine learning for healthcare behavioural OR: Addressing waiting time perceptions in emergency care	2019	4	2,175	94,00
134	The evolution of the Internet of Things (IoT): A computational text analysis	2019	4	2	94,002
135	Big data analytics: Machine learning and Bayesian learning perspectives--What is done? What is not?	2019	3	2,541	93,002541
136	Decision-Making based on Big Data Analytics for People Management in Healthcare Organizations	2019	3	2,415	93,002415
137	Hidden big data analytics issues in the healthcare industry	2019	3	2,297	93,002297
138	Deep Learning versus Conventional Machine Learning for Detection of Healthcare-Associated Infections in French Clinical Narratives	2019	3	1,574	93,00
139	Orthodontics in the era of big data analytics	2019	3	0,946	93,000946
140	Using Big Data Analytics to Advance Precision Radiation Oncology	2018	12	6,203	92,006203
141	Branding luxury hotels: Evidence from the analysis of consumers' "big" visual data on TripAdvisor	2019	2	4,028	92,004
142	A technology delivery system for characterizing the supply side of technology emergence: Illustrated for Big Data & Analytics	2018	12	3,815	92,00

143	Introduction to Machine Learning in Digital Healthcare Epidemiology	2018	12	2,938	92,002938
144	M-Lean: An end-to-end development framework for predictive models in B2B scenarios	2019	2	2,921	92,00
145	Healthcare-associated ventriculitis and meningitis in a neuro-ICU: Incidence and risk factors selected by machine learning approach	2018	12	2,783	92,002783
146	Big data analytics for video surveillance	2019	2	2,101	92,002101
147	Integrating TTF and IDT to evaluate user intention of big data analytics in mobile cloud healthcare system	2019	2	1,429	92,001429
148	Innovative and efficient method of robotics for helping the Parkinson's disease patient using IoT in big data analytics	2019	2	1,258	92,001258
149	Bringing big data analytics closer to practice: A methodological explanation and demonstration of classification algorithms	2019	2	1,225	92,001
150	Determinant factors in applying electronic medical records in healthcare [Facteurs déterminants de l'utilisation de dossiers médicaux électroniques dans les soins de santé]	2019	2	0,678	92,000678
151	Optimizing the electronic health records through big data analytics: A knowledge-based view	2019	1	3,745	91,003745
152	Modern day monitoring and control challenges outlined on an industrial-scale benchmark fermentation process	2019	1	3,334	91,003334
153	Transforming healthcare with big data analytics and artificial intelligence: A systematic mapping study	2019	1	2,95	91,00295
154	Big Data Analytics in Healthcare: Data-Driven Methods for Typical Treatment Pattern Mining	2019	1	1,079	91,001079
155	Hackathons as a means of accelerating scientific discoveries and knowledge transfer	2018	10	9,944	90,009944
156	Efficient development of high performance data analytics in Python	2020	0	5,768	100,005768
157	Tiger hash based AdaBoost machine learning classifier for secured multicasting in mobile healthcare system	2019	0	3,458	90,003458
158	Government initiative in brazilian public health: A technology transfer analysis	2019	0	2,849	90,002849
159	Using big data analytics to improve HIV medical care utilisation in South Carolina: A study protocol	2019	0	2,367	90,002367

160	BioStruct-Africa: Empowering Africa-based scientists through structural biology knowledge transfer and mentoring - Recent advances and future perspectives	2019	0	2,251	90,002
161	Machine Learning-Based Forecast of Hemorrhagic Stroke Healthcare Service Demand considering Air Pollution	2019	0	1,803	90,00
162	Intelligent Classifier: a Tool to Impel Drug Technology Transfer from Academia to Industry	2019	0	1,452	90,001452
163	Pharmacoepidemiology and Big Data Analytics: Challenges and Opportunities when Moving towards Precision Medicine	2019	0	1,096	90,001096
164	Implementation of a knowledge-based manufacturing on the example of sumar tools oü [Teadmuspõhise tootmise juurutamine sumar tools oü-s]	2019	0	0,51	90,00051
165	Using ECHO Clinics to Promote Capacity Building in Clinical Supervision	2018	8	4,435	88,004
166	Energy-efficient acceleration of MapReduce applications using FPGAs	2018	8	1,819	88,001819
167	Real-time big data analytics for hard disk drive predictive maintenance	2018	7	2,189	87,002189
168	A Two-Stage Big Data Analytics Framework with Real World Applications Using Spark Machine Learning and Long Short-Term Memory Network	2018	7	2,143	87,002143
169	Is the eunethta hta core model® fit for purpose? Evaluation from an industry perspective	2018	5	1,494	85,001
170	The Fuzzy System as a Promising Tool for Drugs Selection in Medical Practice	2018	3	3,745	83,004
171	Application of machine learning models in predicting length of stay among healthcare workers in underserved communities in South Africa	2018	3	2,547	83,002547
172	Classification of lung cancer stages with machine learning over big data healthcare framework	2018	1	4,594	81,005
173	Making the Most of Innovation in Personalised Medicine: An EU Strategy for a Faster Bench to Bedside and Beyond Process	2018	0	1,518	80,002

APÊNDICE B – Instrumento de coleta de dados

Variável	Descrição
y	Retornou ou não ao tratamento
x1	Diagnóstico e tratamento anteriores
x2	Ano
x3	Recno - identificador único e sequencial
x4	Tipo de caso
x5	Sexo
x6	Idade
x7	Local do nascimento
x8	Raça / cor
x9	Escolaridade
x10	Clínica do primeiro atendimento
x11	Clínica de início do tratamento
x12	Histórico familiar de câncer
x13	Histórico de consumo de bebida alcoólica
x14	Histórico de consumo de tabaco
x15	Estado de residência
x16	Município de residência
x17	Ano de diagnóstico
x18	Origem do encaminhamento
x19	Exames relevantes para o diagnóstico e planejamento da terapêutica do tumor
x20	Estado civil
x21	Ano da triagem
x22	Ano da primeira consulta
x23	Base mais importante para o diagnóstico do tumor
x24	Localização primária (Categoria 3d)
x25	Localização primária detalhada (Subcategoria 4d)
x26	Tipo histológico do tumor primário
x27	Lateralidade do tumor
x28	Codificação do estádio clínico segundo classificação TNM
x29	Estadiamento clínico do tumor (TNM)
x30	Outros estadiamentos clínicos do tumor
x31	Ptnm - Classificação histopatológica pós-cirúrgica
x32	Principal razão para a não realização do tratamento antineoplásico no hospital
x33	Ano do início do primeiro tratamento específico para o tumor no hospital
x34	Primeiro tratamento recebido no hospital
x35	Estado da doença ao final do primeiro tratamento no hospital
x36	Número do CNES do hospital
x37	UF da unidade hospitalar
x38	Município da unidade hospitalar
x39	Ocupação principal
x40	Data triagem
x41	Data que começou o tratamento

APÊNDICE C – Parecer Consubstanciado CEP

PARECER CONSUBSTANCIADO DO CEP

DADOS DO PROJETO DE PESQUISA

Título da Pesquisa: Análise big data e transferência de tecnologia: uma ferramenta para análise preditiva aplicada aos cuidados de saúde

Pesquisador: JOAO LUIZ KOVALESKI

Área Temática:

Versão: 2

CAAE: 50596221.7.0000.5547

Instituição Proponente: UNIVERSIDADE TECNOLOGICA FEDERAL DO PARANA

Patrocinador Principal: Financiamento Próprio

DADOS DO PARECER

Número do Parecer: 4.993.564

Apresentação do Projeto:

De acordo com o pesquisador:

Desenvolvida a partir do conceito de inteligência de negócios e sistemas de suporte a decisão, a análise big data representa um conjunto de técnicas, tecnologias, sistemas, práticas, metodologias e aplicações com capacidade analítica para grandes conjuntos de dados onde procuram esclarecer o que está obscuro e subentendido para as organizações por meio da descoberta de informações sobre negócios, mercado e tomada de decisão (AMALINA et al., 2019; MOHAMED et al., 2020).Na era digital, o grande volume de dados e informações que está sendo gerado em velocidades e variedades cada vez maiores, expressam a complexidade da análise big data enquanto prática organizacional (WANG; KUNG; BYRD, 2018). Outro assim, enfrentamentos da implementação de arquiteturas de análise big data e a remodelagem da capacidade tecnológica colocam organizações de saúde em situação sensível para obtenção de benefícios, os quais transformam a prática clínica (GALETSI; KATSALIAKI; KUMAR,2020). Uma alternativa para obter-se a capacidade tecnológica necessária a análise big data está no desenvolvimento de novos aplicativos com recursos big data, os profissionais de tecnologia de informação estão dispostos a contribuir com as partes interessadas para aumentar o poder analítico das organizações intensivas em dados (PANAYIDES et al., 2018; RAZZAK; IMRAN; XU, 2020). Nesse sentido, as técnicas analíticas sofreram significativas mudanças. Recentemente o aprendizado de máquina, uma abordagem computacional orientada a dados baseada na utilização de

algoritmos capazes de fazer previsões e predições de resultados são peças essenciais neste contexto (MCNUTT et al., 2018).As proximidades entre análise big data e aprendizado de máquina desenvolve grande potencial para ser utilizado de forma sistemática em conjuntos de dados, procurando descobrir padrões interessantes que até então eram desconhecidos e ainda, descobrir ineficiências nos armazenamentos de dados a fim de desenvolver modelos preditivos objetivando melhorar a qualidade da saúde e redução de custos. Contudo, com todas as suas funcionalidades estabelecidas, os processos de transferência de tecnologia aparecem como medida intrínseca ao resultado da análise (GUO; CHEN, 2019; PAN et al., 2019; PASSOS et al., 2019).Não somente demonstrar os resultados através da visualização de dados, mas tornar acessível aos profissionais envolvidos no setor de saúde as etapas pregressas através de procedimentos preestabelecidos, etapas de diagnóstico, tratamento e acompanhamento dependem da aquisição e interpretação dos dados, mesmo com as melhorias substanciais recebidas nos últimos anos, ainda sofrem dificuldades em virtude das características do big data incluindo volume, velocidade, variedade, variabilidade, veracidade, valor e valência (SAGGI; JAIN, 2018). Com esta complexidade construída pela diversidade de fonte de dados, novas técnicas são requeridas para lidar com os conjuntos de dados expressivos (WANG; HAJLI, 2017; SHAFQAT et al., 2020). A coexistência de técnicas analíticas avançadas está na combinação das linguagens de programação como Python, Scala, R, SQL (JONES et al., 2018; MOHAMED et al., 2020).Neste contexto, a aplicação do algoritmo desenvolvendo capacidade analítica apropriada a situação de investigação irá produzir resultados dos quais necessitam ser submetidos a medidas de desempenho para sua validação (WARING; LINDVALL; UMETON, 2020). Profissionais de saúde necessitam de familiaridade com as técnicas computacionais, justamente para estruturar o processo de tomada de decisão e transferência de tecnologia é preciso evitar a tomada de decisão algorítmica, devido a real necessidade de avaliação e interpretação sobre os resultados apresentados por partes dos profissionais envolvidos. Ainda assim, as medidas de transferência de tecnologia precisam objetivar a significância clínica e a redução de custo voltadas ao tratamento do paciente, com monitoramento constante e melhoria de desempenho (MILLER; FRENCH, 2016; PAN et al., 2019; HAQ et al., 2020). Diante das evidências descritas, esta pesquisa desenvolveu o seguinte problema de pesquisa: Como a análise big data pode contribuir para a gestão e transferência de tecnologia explorando grandes conjuntos de dados reais de pacientes acometidos por
doença, visando a predição da piora em nível clínico?
Hipótese:
Ferramenta desenvolve capacidade preditiva de mensurar a taxa de piora clínica de pacientes

submetidos
aos cuidados de saúde, orientando a tomada de decisão clínica e transferência de tecnologia com bases em dados que não receberam tratamento analítico.
Metodologia Proposta:
Sobre a ótica para abordar a natureza da pesquisa, a classificação aplicada é apropriada devido ao seu desenvolvimento em uma determinada realidade, objetivando contribuir com processos e conhecimentos produzidos em uma situação real (GIL, 1999).Na abordagem do problema, a classificação é qualitativa quantitativa.
Para Creswell (2010) esta abordagem se define pela aplicação dos pontos fortes tanto dos métodos qualitativos quanto quantitativos, possibilitando maior compreensão sobre o problema estudado. A cerca dos objetivos, esta pesquisa é classificada como exploratória, por esclarecer e propor uma aproximação com o problema de estudo, conhecendo de forma detalhada as variáveis do estudo, explorando métodos e técnicas aplicadas em um fenômeno desconhecido através de levantamento de experiências, observação informal e fonte de dados secundários (LAKATOS; MARCONI, 2001).Quanto aos procedimentos técnicos, a pesquisa é estabelecida como experimental e documental devido a aplicação de técnicas estatísticas visando a predição da piora em nível clínico através de uma ferramenta, além de contribuir com a compreensão do que estão inseridos em bancos de dados dos quais não receberam algum tipo de tratamento analítico. Para a pesquisa experimental, Gil (1999) esclarece que o delineamento está na determinação de um objeto de estudo, na seleção das variáveis que seriam capazes de influenciá-lo, estabelecendo formas de controle e de observação dos efeitos que a variável produz. Aliado a isto, a utilização de técnicas estatísticas é predominantemente aceita. A pesquisa documental possui duas estratégias de coleta de dados, sendo o local onde os documentos são encontrados e coletados, possibilitando dois ambientes, o campo ou o laboratório, e a segunda estratégia refere-se a fonte dos dados podendo estar nos documentos ou no campo (APPOLINÁRIO, 2009).Primeira Etapa: Revisão Sistemática de Literatura Segunda Etapa: Desenvolvimento da Ferramenta de Análise Big data e TT Orientada a Análise Preditiva. Para compreender a solução proposta para o problema constituído nesta pesquisa, foi desenvolvida uma ferramenta baseada na arquitetura de análise big data e TT em virtude da proporção de dados e informações em um formato contínuo e intenso acerca de pacientes submetidos aos cuidados de saúde, composto por cinco estágios inter-

relacionados (Figura 04). No estágio 1, denominado captura de dados, considera diversas naturezas de dados, por exemplo: dados clínicos, exames, prontuários eletrônicos médicos, prescrições diversas, dados de identificação, dentre outros. Terceira Etapa: aplicação das medidas de desempenho para validação do algoritmo regressão logística. Diferentes métricas são aplicadas para a medição de desempenho da regressão logística, acurácia, especificidade, sensibilidade, recall, precisão, pontuação F1, curva ROC,AUC. A matriz de confusão binária foi desenvolvida a fim de calcular essas medidas de desempenho (JOHNSTON et al., 2019; HAQ et al., 2020; LI et al., 2020).Nesta pesquisa, considerou-se os dados que se encontram no banco de dados da instituição hospitalar, dados do prontuário eletrônico, dados de exames laboratoriais, de imagens e dados inseridos no RHC. A forma de apresentação é através de relatórios e planilhas. Os dados são de um conjunto de pacientes acometidos por câncer de mama da qual necessita de acompanhamento médico para o tratamento, com base nos critérios de seleção e preditores preestabelecidos. Para realização do estágio 2 da ferramenta será utilizado o pacote Dplyr no software Rstudio para efetuar a limpeza, transformação e seleção dos dados. Para desenvolver o estágio 3 “aplicação do algoritmo” e o estágio 4 “descoberta do conhecimento” será utilizado o pacote de predição do nível do paciente do Programa de Ciência e Informática em Saúde Observacional (OHDSI).

Critério de Inclusão:
1ª Critério: pacientes com idade igual ou maior de 18 anos;2ª Critério: Submetidos ao primeiro tratamento em unidade hospitalar entre 2009-2019;3ª Critério: Localização primária identificada;4ª Critério: Com estágio da doença em progressão ao término do primeiro tratamento.
Critério de Exclusão:
1ª Critério: Pacientes que desistiram ou não completaram do tratamento recomendado;
2ª Critério: Pacientes que receberam o primeiro tratamento por doença não relacionada ao câncer de mama;
3ª Critério: Pacientes não encaminhados ao primeiro tratamento através do Sistema Único de Saúde (SUS);
4ª Critério: Pacientes com doenças preexistentes.

Objetivo da Pesquisa:

De acordo com o pesquisador:

Objetivo Primário:

Desenvolver uma ferramenta com funcionalidades de análise big data e transferência de tecnologia para preditar os resultados de pacientes submetidos a cuidados de saúde.

Objetivo Secundário:

OE 1. Determinar as principais características da análise big data e Transferência de Tecnologia (TT) no contexto de cuidados de saúde;

OE 2.Apresentar formas de aplicação de aprendizado de máquina voltada a análise de grandes conjuntos de dados;

OE 3. Produzir uma ferramenta com funcionalidades de análise big data e TT apresentando o sistema de aprendizagem em saúde orientado a análises preditivas;

OE 4.Analisar o desempenho da ferramenta através de um conjunto de dados reais de pacientes acometidos por doença por meio das medidas de desempenho.

Avaliação dos Riscos e Benefícios:

De acordo com o pesquisador:

Riscos:

Os riscos que podem ser encontrados no ambiente de pesquisa são:•Dificuldades de acesso aos dados nas bases de dados disponibilizada, tendo em vista que o estudo retrospectivo se desenvolve no período de 2009-2019;•Não produção de relatórios por parte do sistema de registros hospitalares de câncer ou sistema de gerenciamento de prontuário eletrônico;•Incorrência com os preditores e os dados inseridos nas bases de dados disponibilizadas;•Riscos biológicos e ergonômicos por parte do membro da equipe de pesquisa envolvido diretamente com o ambiente de

coleta de dados.

Benefícios:

Os benefícios que podem ser desenvolvidos no ambiente de pesquisa são:•Possibilidade de exploração dos bancos de dados a partir de uma ferramenta de análise preditiva que encontra a taxa de piora clínica em pacientes em tratamento;•Colaboração na formação de nova habilidade advinda da ciência de dados a partir da experiência desenvolvida de forma conjunta;•Disponibilização a ferramenta produzida e validada a instituição coparticipante;•Contribuição à equipe médica no processo decisório na manutenção de equipes de enfermagem e multiprofissionais, amparado pelos resultados produzidos pela ferramenta no atendimento da

piora clínica em unidade hospitalar.

Comentários e Considerações sobre a Pesquisa:

O projeto atende as recomendações da Resolução 466/2012.

Considerações sobre os Termos de apresentação obrigatória:

A pesquisa pretende desenvolver uma ferramenta com funcionalidades de análise big data e transferência de tecnologia para preditar os resultados de pacientes submetidos a cuidados de saúde.

Recomendações:

No último parecer consubstanciado foram elencadas as seguintes recomendações:

1. Anexar o termo de confidencialidade de dados e entrega de relatório final contendo as assinaturas de todos os membros da equipe de pesquisa (falta a assinatura da profa. Regina Negri Pagani).
ATENDIDO
2. Rever os critérios de inclusão e exclusão, uma vez que são o oposto um do outro. Mesmo sendo um estudo retrospectivo, devem ser definidos.
ATENDIDO
3. Rever o item de intervenções a serem realizadas no formulário da plataforma Brasil. Consta "não se aplica". No entanto, mesmo sendo análise de prontuários, devem ser especificadas todas as etapas e informações necessárias.
ATENDIDO
4. Rever cronograma em função do calendário de reuniões do CEP.
ATENDIDO

Conclusões ou Pendências e Lista de Inadequações:

Todas as recomendações solicitadas foram atendidas.

Considerações Finais a critério do CEP:

Diante do exposto, o CEP-UTFPR, de acordo com as atribuições definidas no cumprimento da Resolução CNS nº 466 de 2012, Resolução CNS nº 510 de 2016 e da Norma Operacional nº 001 de 2013 do CNS, manifesta-se por APROVAR este projeto.

Lembramos aos (as) senhores(as) pesquisadores(as) que o Comitê de Ética em Pesquisa (CEP) deverá receber relatórios anuais sobre o andamento do estudo, bem como a qualquer tempo e a critério do pesquisador nos casos de relevância, além do envio dos relatos de eventos adversos,

para conhecimento deste Comitê. Salientamos ainda, a necessidade de relatório completo ao final do estudo. Eventuais modificações ou emendas ao protocolo devem ser apresentadas ao CEP-UTFPR de forma clara e sucinta, identificando a parte do protocolo a ser modificado e as suas justificativas.

Este parecer foi elaborado baseado nos documentos abaixo relacionados:

Tipo Documento	Arquivo	Postagem	Autor	Situação
Informações Básicas do Projeto	PB_INFORMAÇÕES_BÁSICAS_DO_PROJETO_1800910.pdf	30/08/2021 18:40:32		Aceito
Projeto Detalhado / Brochura Investigador	modelo_projeto_CEP.pdf	30/08/2021 18:38:26	JOAO LUIZ KOVALESKI	Aceito
Declaração de Pesquisadores	Termo_de_compromisso_de_confidencialidade_de_dados_completo.pdf	30/08/2021 18:25:14	JOAO LUIZ KOVALESKI	Aceito
Folha de Rosto	Folha_de_Rosto_assinada.pdf	09/08/2021 14:02:25	JOAO LUIZ KOVALESKI	Aceito
Outros	Instrumento_de_coleta_de_dados.pdf	29/07/2021 15:22:10	JOAO LUIZ KOVALESKI	Aceito
Outros	TCUD.pdf	29/07/2021 15:21:13	JOAO LUIZ KOVALESKI	Aceito
Declaração de concordância	carta_de_aceite_do_projeto_de_pesquisa.pdf	29/07/2021 15:20:39	JOAO LUIZ KOVALESKI	Aceito
TCLE / Termos de Assentimento / Justificativa de Ausência	Solicitacao_Dispensa_TCLE.pdf	29/07/2021 15:18:56	JOAO LUIZ KOVALESKI	Aceito

Situação do Parecer:
Aprovado

Necessita Apreciação da CONEP:
Não

CURITIBA, 23 de Setembro de 2021

Assinado por:
Frieda Saicla Barros
(Coordenador(a))

Continuação do Parecer: 4.993.564

Endereço: SETE DE SETEMBRO 3165
Bairro: CENTRO **CEP:** 80.230-901
UF: PR **Município:** CURITIBA
Telefone: (41)3310-4494 **E-mail:** coep@utfpr.edu.br

SANTA CASA DE PONTA GROSSA	DOCUMENTO DE GESTÃO	Revisão	003
		Data	05/11/2015
	DG - 481	Página	01 de 01

Carta de Aceite de Projeto de Pesquisa

Ponta Grossa, 28 de Junho de 2021.

Prezados pesquisadores Myller Augusto Santos Gomes, após avaliação do seu projeto em pesquisa intitulado por **"Análise Big Data e transferência de tecnologia: Uma ferramenta para análise preditiva aplicada a cuidados de saúde",** informamos que esta comissão aceita o seu desenvolvimento na Santa Casa de Misericórdia de Ponta Grossa, sendo que o início do mesmo será permitido após a apresentação do documento referente a aprovação do comitê de ética em pesquisa a qual seu projeto foi submetido.

Atenciosamente,

		Fernanda Rachel C. da Silva Gerente Contratos e Serviços
Dr. Rogerio Santos Clemente *Diretor Técnico* *CRM 6934*	**Dr. Marcelo Derbli Schafranski** *Presidente da Comissão de Avaliação em Pesquisa - COAP* *CRM 17192*	**Fernanda Rachel Camargo da Silva** *Gerente de contrato e serviços - SCMPG*