UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ

IGOR AUGUSTO SANTOS

UTILIZAÇÃO DO BANCO DE DADOS INTERNACIONAL DE TESTES DE ACIDENTE VASCULAR CEREBRAL PARA APLICAÇÃO DA METODOLOGIA *KNOWLEDGE DISCOVERY IN DATABASES*

LONDRINA

2021

IGOR AUGUSTO SANTOS

UTILIZAÇÃO DO BANCO DE DADOS INTERNACIONAL DE TESTES DE ACIDENTE VASCULAR CEREBRAL PARA APLICAÇÃO DA METODOLOGIA *KNOWLEDGE DISCOVERY IN DATABASES*

Trabalho de conclusão de curso de graduação apresentada como requisito para obtenção do título de Bacharel em Engenharia de Produção da Universidade Tecnológica Federal do Paraná (UTFPR).
Orientador(a): Prof. Dr. Bruno Samways dos Santos

LONDRINA

2021

IGOR AUGUSTO SANTOS

UTILIZAÇÃO DO BANCO DE DADOS INTERNACIONAL DE TESTES DE ACIDENTE VASCULAR CEREBRAL PARA APLICAÇÃO DA METODOLOGIA *KNOWLEDGE DISCOVERY IN DATABASES*

Trabalho de Conclusão de Curso de Graduação para obtenção do título de Bacharel em Engenharia De Produção da Universidade Tecnológica Federal do Paraná (UTFPR).

Data de aprovação: 02 de Dezembro de 2021

_____

Bruno Samways dos Santos
Doutor
Universidade Tecnológica Federal do Paraná

_____

Rogério Tondato
Doutor
Universidade Tecnológica Federal do Paraná

_____

Pedro Rochavetz de Lara Andrade
Doutor
Universidade Tecnológica Federal do Paraná

LONDRINA

2021

AGRADECIMENTOS

O desafio e a coragem de começar a segunda graduação foram alicerçados pelo apoio, suporte, conselhos e orientações de várias pessoas do período que morava em São Paulo. Nesse momento, registro minha gratidão para essas pessoas em nome da Professora Dra. Mirian Rejowski da Universidade Anhembi Morumbi, campus Vila Olimpia, foi a primeira pessoa a saber da minha aprovação na UTFPR, conversamos, me orientou, transmitiu saberes, dando suporte e apoio incondicional nesse recomeço. Muito obrigado professora!

Ao sair de São Paulo e ir para Medianeira, oeste paranaense, tive a oportunidade de encontrar a paz e a calmaria. Durante esse período várias amizades foram feitas, nos tornamos como uma grande família. Aos meus amigos de Medianeira, o meu muito obrigado!

Com a oportunidade de ficar mais próximo de São Paulo, transferi para o campus da UTFPR de Londrina, um novo recomeço. Nova cidade, novas amizades e a oportunidade de ingressar no mercado de trabalho na área de engenharia de produção. Aos meus amigos de Londrina, o meu muito obrigado!

Todos os discentes que apoiou, participou e marcou presença nas atividades, visitas, cursos e palestras organizadas pela Liga Universitária Integra Engenharia, da qual fui um dos membros fundadores, cuja proposta foi a de estreitar a relação entre academia e mercado de trabalho. Meu muito obrigado!

Ao longo dessa jornada, vários docentes fizeram parte da minha formação, e é o momento de registrar gratidão a todos que marcaram e contribuíram para o meu desenvolvimento.

Aos professores Dr. Alireza Mohebi Ashtiani e professora Dra. Nazira Hanna Harb, ambos do departamento de matemática, muito obrigado pelos conselhos, orientações e por toda paciência nas aulas de cálculo diferencial e equações diferenciais ordinárias. Parabenizá-los pela didática, pela acessibilidade, pela prestatividade, pela empatia e humanismo que possuem. Vocês são incríveis. Muito obrigado!

Mesmo com a incerteza e insegurança, a senhora me deu a oportunidade de ser seu monitor. O que seria inicialmente um teste de seis meses, se tornaram 18 meses. Professora Dra. Marilene Turini Piccinato do departamento de Física, o meu muito obrigado pela oportunidade de trabalharmos juntos, pelos diálogos e pela transparência durante esse período. Obrigado pela oportunidade e pela confiança em ser seu monitor de Física 2.

Aos professores Dr. Raphael Euclides Prestes Salem e professor Dr. Ricardo de Vasconcelos Salvo com quem tive a oportunidade de ser discente das disciplinas de mecânica geral 2 e mecânica geral 1, respectivamente. Pela acessibilidade nos atendimentos, pela paciência em tirar dúvidas, pelos incentivos em perseverar até o final do semestre, muito obrigado!

Em nome do DAENP, minha gratidão ao professor Dr. Rogério Tondato, em suas aulas, sempre prestativo e atencioso. Me aconselhou em várias situações, e no findar desse ciclo se fez presente na minha banca, me acompanhou do começo ao fim dessa jornada. Muito obrigado professor!

Com o senhor tive o primeiro contato com a logística em sala de aula, área que se tornaria anos mais tarde o meu caminho de atuação profissional, depois, disciplinas de sistemas produtivos e gestão de projetos. Várias disciplinas, muitos diálogos, muitos saberes. Ao professor Dr. Edilson Giffhorn por tudo que me ensinou, por todos os conselhos e orientações, minha admiração e gratidão por sua contribuição durante minha formação. Muito obrigado professor!

A inspiração desse trabalho ocorreu em virtude do falecimento do meu pai no ano de 2017, ali surgiu o interesse em fazer o trabalho de conclusão de curso na área de saúde. Ao professor Dr. Bruno Samways dos Santos minha gratidão pela orientação, pelos conselhos, paciência, disposição, por tudo que colaborou e ensinou. Por ter aceitado ser meu orientador, por ter atuado como um mentor durante a graduação. Muito obrigado professor!

Todas as pessoas citadas anteriormente possuem grande valor e significado na minha jornada, mas sem o apoio incondicional e suporte dos meus pais, esse caminho não seria possível. Aos meus pais, o meu muito obrigado por tudo! Esse ciclo que se encerra é fruto de tudo que me ensinaram e apoiaram, esse trabalho é para vocês! Minha eterna gratidão!

RESUMO

De acordo com o Plano de Ações Estratégicas para o Enfrentamento das Doenças Crônicas Não Transmissíveis (DCNT) no Brasil 2011-2022 (2011), as DCNTs constituem o problema de saúde de maior relevância e respondem por mais de 70% das causas de mortes no Brasil. Considerando a amplitude da saúde em sua forma macro, o presente trabalho fez um recorte sobre o acidente vascular cerebral (AVC), com o objetivo de classificar pacientes que possam evoluir para óbito em até seis meses por meio de técnicas de mineração de dados. Partindo desse pressuposto, e utilizando como referência o Banco Internacional de Testes de AVC, o qual, contempla 3034 instâncias e 266 atributos, foi feito identificação de trabalhos correlatos que por sua vez atuaram como suporte de sustentação na escolha e definição de atributos com o intuito de predizer pacientes diagnosticados com AVC que podem evoluir para óbito. Quatro trabalhos foram norteadores da escolha dos atributos, aplicando-se as técnicas de árvore de decisão e floresta aleatória de tal modo que, foram aplicadas as técnicas para quatro conjunto de dados distintos. Dessa forma, foi possível mensurar a performance de classificação, bem como avaliar os atributos que melhor se encaixam para predição proposta. O melhor resultado foi o que teve como referência os atributos utilizados no estudo de Dheepitha Babu *et.,al* (2021), sendo 69,67 de acurácia, precisão de 0,749, seguido pela sensibilidade de 0,794. Os estudos selecionados tiveram papel de dar sustentação no processo de seleção dos atributos, por outro lado, a sensibilidade da área da saúde demanda por melhores resultados. Assim, é recomendável o uso da base utilizada no presente trabalho, bem como das métricas mencionadas, com a ressalva de ter como referências outros atributos e técnicas, objetivando melhores resultados.

Palavras-chave: AVC; KDD; mineração de dados; classificação.

ABSTRACT

According to the Strategic Action Plan for Confronting Chronic Non-Communicable Diseases (NCDs) in Brazil 2011-2022 (2011), as previous CNCDs, the health problem of greater derivation and due to more than 70% of the causes of deaths in Brazil. Brazil. Observing the breadth of health in its macro form, the present work made a cut about the cerebrovascular accident (CVA), with the objective of classifying the patients that can evolve to death in up to six months through data mining techniques. Based on this assumption, and taking as a reference the International Bank of CVA Tests, which includes 3034 computed and 266 attributes, correlated works were identified, which in turn acted as supporting support in the choice and definition of attributes with the purpose to predict patients diagnosed with stroke that may progress to death. Four works guided the choice of attributes, applying them as decision tree and random forest techniques in such a way that they were applied as techniques for four distinct data sets. In this way, it was possible to measure the classification performance, as well as evaluate the attributes that best fit the proposed prediction. The best result was the one that had as reference the attributes used in the study by Dheepitha Babu et., Al (2021), with 69.67 of accuracy, precision of 0.749, followed by sensitivity of 0.794. The selected studies had the role of supporting the attribute selection process, on the other hand, a sensitivity in the health area demands better results. Thus, it is recommended to use the base used in this work, as well as the mentioned metrics, with the exception of having references to other attribute and techniques, aiming at better results.

Keywords: stroke; KDD; data mining; classification.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

SUMÁRIO

1. INTRODUÇÃO

Os dados são originados de indústrias dos mais diversos ramos de produção, empresas de telecomunicações, instituições educacionais, hospitais, instituições financeiras, de saúde, dentre tantas outras. No entanto, a tarefa de simplesmente armazenar tais dados não é suficiente, é necessário, também, verificar se os dados coletados possuem informações relevantes e se há algum conhecimento a ser descoberto. Segundo Góes e Steiner (2012), a geração de banco de dados ocorre naturalmente atualmente, pois os meios computacionais são práticos para seu armazenamento.

Desta forma, a identificação dessas informações e possíveis descobertas serão obtidas a partir da aplicação de técnicas existentes, tais como: *data mining e machine learning. Data mining* (DM), também conhecida como mineração de dados e, segundo Ragsdale (2014) é o processo de descoberta e extração de informações e conhecimentos não triviais de grandes conjuntos de dados. *Machine learning* (ML), que traduzido para o português corresponde ao aprendizado de máquina que, de acordo com Mitchell (1997), é definido como a capacidade de melhorar o desempenho na realização de alguma tarefa por meio da experiência.

Em complemento à essa definição, Faceli *et al.* (2011), destaca aplicações bem sucedidas de técnicas de ML na solução de problemas reais, tais como: reconhecimento de palavras faladas, predição de taxas de cura de pacientes com diferentes doenças, detecção no uso fraudulento de cartões de crédito, condução de automóveis de forma autônoma em rodovias, ferramentas que jogam gamão e xadrez de forma semelhante a campeões, diagnóstico de câncer por meio da análise de dados de expressão gênica, entre outros.

Assim, a saúde face a infinidade de problemas e desafios demandando por soluções torna-se palco de oportunidades para aplicações de técnicas existentes. De um lado, uma área que lida com várias situações que requerem distintas abordagens e tratamentos, do outro, a área que contempla técnicas voltadas ao tratamento de grandes conjuntos de dados, cuja proposta é a de identificar padrões que possam contribuir significativamente na elaboração de estratégias a fim de encontrar boas soluções dos problemas alvo da pesquisa.

De acordo com a Organização Mundial de Saúde (2018), estima-se que 13 milhões de pessoas morrem todos os anos antes dos 70 anos por doenças cardiovasculares, doenças respiratórias crônicas, diabetes e câncer – a maioria delas em países de baixa e média renda, e que, em 2016, morreram por dia 15 mil crianças menores de cinco anos. De acordo com o Plano de Ações Estratégicas para o Enfrentamento das Doenças Crônicas Não Transmissíveis (DCNT) no Brasil 2011-2022 (2011), as DCNTs constituem o problema de saúde de maior relevância e respondem por mais de 70% das causas de mortes no Brasil. Neste contexto, de acordo com a Secretaria Estadual de Saúde do Paraná no período de 2008 a 2017 houveram 26.825 registros de óbitos apenas de doenças cardiovasculares, sendo que deste número, 9.434 óbitos foram identificados apenas no município de Londrina. A Secretaria estadual destaca que: esse conjunto de doenças correspondeu a 59% de todas as mortes e 43% desses óbitos ocorreram na faixa etária de 30 a 69 anos.

De um lado, o cenário onde a OMS apresenta uma meta até 2030, do outro, países como o Brasil com o desafio de desenvolver e buscar alternativas para contribuir com a redução de mortalidade prematura de doenças crônicas não transmissíveis. Desse modo, o presente trabalho, através da utilização das técnicas de classificação, árvore de decisão e floresta aleatória irá aplicar as referidas técnicas no Banco de Dados Internacional de Testes de Acidente Vascular Cerebral.

Neste sentido, a proposta deste trabalho é a de responder a seguinte pergunta de partida: Como a metodologia KDD pode auxiliar na predição da evolução para óbito em pacientes diagnosticados com AVC?

1.1 Objetivo Geral

Classificar a evolução de óbitos de pacientes diagnosticados com AVC por meio da metodologia KDD aplicada em banco de dados internacional de testes de AVC.

1.2 Objetivos Específicos

- Compor o estado da arte sobre tarefas de classificação quanto ao diagnóstico de AVC.
- Testar o modelo de classificação em instâncias reduzidas a partir de atributos indicados pela literatura.
- Apoiar a meta: 3:41 estabelecida pela OMS: “Até 2030, reduzir um terço de mortalidade prematura por doenças crônicas não transmissíveis via prevenção e tratamento e promover a saúde mental e o bem-estar especificado pela OMS.

1.3 Justificativa

Em maio de 2018 foi publicado no site da OMS, as estatísticas da saúde, sendo visualizado a dimensão dos vários problemas de saúde a nível global. Diante a vasta extensão de problemas, a matéria sobre as estatísticas da saúde aborda os Objetivos Sustentável (ODS), de tal modo que, dentre os ODS’s listados, foi escolhido o ODS 3 intitulado: Assegurar uma vida saudável e promover o bem estar para todos, em todas as idades, que por sua vez é composto por várias metas onde a meta escolhida e norteadora desse trabalho é a denominada: 3.41, intitulada: “Até 2030, reduzir um terço a mortalidade prematura por doenças crônicas não transmissíveis via prevenção e tratamento, e promover a saúde mental e o bem-estar”.

A Organização Mundial da Saúde (OMS) define como doenças crônicas as doenças cardiovasculares (cerebrovasculares, isquêmicas), as neoplasias, as doenças respiratórias crônicas e diabetes *mellitus*. Considerando esse rol de doenças, foi selecionado acidente vascular cerebral que faz parte das doenças cardiovasculares para temática e aplicação da metodologia KDD, metodologia a qual está em consonância com a proposta do presente trabalho, descobrir conhecimento a partir da extração em banco de dados. Essa metodologia está alicerçada nas seguintes fases: seleção de dados, limpeza dos dados ou pré-processamento, transformação dos dados, *data mining*, interpretação do conhecimento gerado.

Assim, a primeira fase se enquadra na seleção de dados do público analisado, passando por limpeza desses dados, seguido pela aplicação de técnicas analíticas e por fim interpretação das informações obtidas.

Como consequência, tenta-se identificar características, padrões e fatores que possam predizer a evolução para óbito de pacientes diagnosticados com a acidente vascular cerebral.

A meta 3.41 tem alcance a nível global, ou seja, alcança todos os países dentre eles o Brasil. Porém, por sua dimensão continental e as distintas variáveis a serem consideradas em cada região, reduziu-se a aplicação do presente trabalho em direcionar os esforços e aplicação da metodologia KDD no banco de dados internacional de testes de acidente vascular cerebral.

No geral, as doenças cardíacas continuaram sendo uma das maiores causas de morte nas últimas décadas, e é por esse motivo que as doenças cardíacas são consideradas uma das principais prioridades da pesquisa médica, que por sua vez gera enormes quantidade de dados. Esses volumes crescentes de dados são muitos adequados para serem processados com técnicas de DM que podem lidar com esses dados com eficiências.

1.4 Classificação da pesquisa e estrutura do trabalho

Em relação a natureza do método da pesquisa, será quantitativa. Segundo Richardson (1999), a pesquisa quantitativa é caracterizada pelo emprego da quantificação, tanto nas modalidades de coleta de informações quanto no tratamento delas por meio de técnicas estatísticas. No presente trabalho, o tratamento dos dados obtidos foi realizado com a metodologia KDD. Seguindo Appolinario (2016, p.22) afirma que “as pesquisas quantitativas seriam aquelas que lidariam como os fatos (característicos nas ciências naturais)”.

O método de pesquisa é modelagem, que por sua vez de acordo com Longaray (2013) de modo geral, pode-se dizer que o processo de modelagem consiste em um conjunto de procedimentos adotados para construir um esquema que represente o problema. Nesse processo, deve-se levar em conta que o mundo é formado por eventos dinâmicos e que tudo o que nos cerca não segue, necessariamente, uma logica pré-determinada

Em relação ao objetivo da pesquisa, possui caráter descritivo, que de acordo com Triviños (1987, p.110), “o estudo descritivo pretende descrever com exatidão os fatos e fenômenos de determinada realidade”.

Assim, o trabalho se divide em 5 seções, organizadas na forma a seguir. A seção 1 compreendeu uma breve contextualização sobre o tema abordado: os objetivos geral e específicos, justificativa e contribuição do trabalho. A seção 2 descreve os problemas cardiovasculares, seguido pelo processo do KDD, técnicas e métricas de avaliação para os problemas de classificação, com os trabalhos correlatos na sequência. A seção 3 sumariza e descreve o conjunto de dados utilizado na pesquisa, discorre também sobre a metodologia utilizada, mostrando os passos aplicados para a execução da pesquisa, bem como a ferramenta (*software*) utilizado para a implementação dos modelos. A seção 4 mostra os resultados encontrados com os algoritmos de classificação. Com a conclusão do estudo sendo descrita na seção 5.

2. FUNDAMENTAÇÃO TEORICA

2.1 Acidente Vascular Cerebral

Segundo Segundo Marchioni e Fisberg (2009, p.1) as DCNT são representadas por um grupo de doenças caracterizadas por história natural prolongada, multiplicidade de fatores de risco complexos, interação de fatores etiológicos conhecidos e desconhecidos, extenso período de latência, longo curso assintomático, curso clínico em geral lento, prolongado e permanente, manifestações clínicas com períodos de remissão e exacerbação e evolução para graus variados de incapacidade ou morte. O grupo das DCNT compreende, majoritariamente, doenças cardiovasculares, diabetes, câncer e doenças crônicas. Contudo, o foco deste trabalho será direcionado às doenças cardiovasculares, que por sua vez se trata de um grupo de doenças do coração e vaso sanguíneos que conforme apresentado pela OMS (2017), incluem:

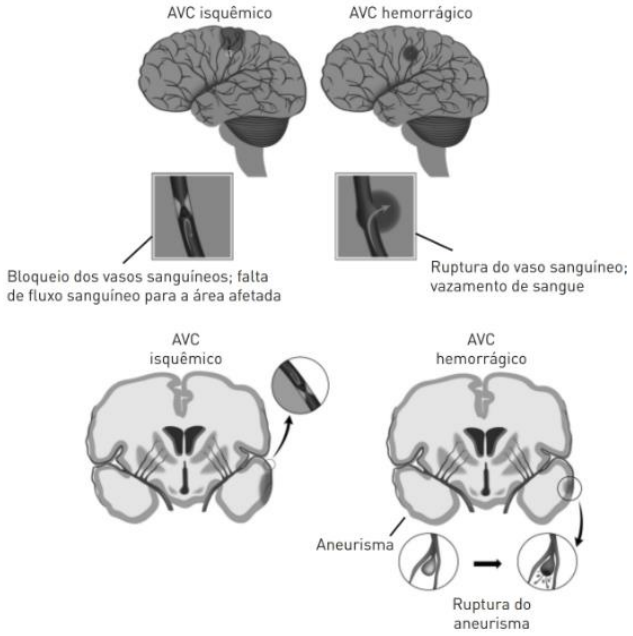
Doença coronariana – doença dos vasos sanguíneos que irrigam o músculo cardíaco. Doença cerebrovascular – doença dos vasos sanguíneos que irrigam o cérebro. Doença arterial periférica – doença dos vasos sanguíneos que irrigam os membros superiores e inferiores. Doença cardíaca reumática – danos do músculo do

coração e válvulas cardíacas devido à febre reumática, causada por bactérias estreptocócicas. Cardiopatia congênita – malformações na estrutura do coração existentes desde o momento do nascimento. Trombose venosa profunda e embolia pulmonar – coágulos sanguíneos nas veias das pernas, que podem se desalojar e se mover para o coração pulmões. (OMS, 2017).

É importante, acrescentar à abordagem da OMS que segundo Jaeger e Manenti (2014) que doença coronariana ocorre quando uma artéria coronária (que fornece o suprimento sanguíneo ao músculo cardíaco) é obstruída, a parte do coração por ela irrigada sofre isquemia, que é a privação de energia (glicose e oxigênio) para esse segmento de músculo. Quando a isquemia se prolonga, as células do coração não conseguem manter sua integridade funcional, perdem a capacidade de realizar o trabalho de contração, posteriormente tornam-se estruturas inviáveis e morrem.

Dentre as doenças cerebrovasculares existentes, Friederich (2014) destaca que o acidente vascular cerebral (AVC) acontece quando um vaso sanguíneo que está nutrindo uma região do cérebro é obstruído por um coágulo de sangue, chamado de isquêmico (AVC isquêmico), ou quando esse vaso se rompe – nesse caso chamado de hemorrágico (AVC hemorrágico) –, causando prejuízo aos neurônios e vias neuronais cerebrais, levando a sintomas e sinais neurológicos conforme a área afetada (Figura 1).

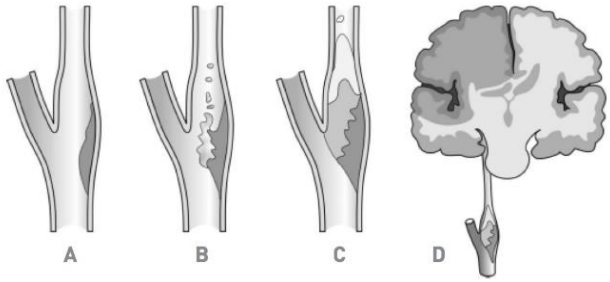
Figura 1: Tipos de AVC.



Fonte: Friederich (2014).

Para doença arterial periférica é relevante ponderar a abordagem de Saadi (2014): As artérias carótidas levam sangue ao cérebro. A maioria dos pacientes com estenose (estreitamento) de carótida não apresenta sintomas, já́ que essas lesões se desenvolvem lentamente ao longo de décadas. Quando ocorre estreitamento por placas de ateroma, em geral na bifurcação da artéria no pescoço, o risco é a liberação de pequenos fragmentos com prejuízo à circulação cerebral conforme indicado na Figura 2.

Figura 2: Estreitamento Gradual da artéria carótida



Fonte: Saadi (2014).

Se a bifurcação na artéria do pescoço for atingida, aos poucos o ateroma aumenta de tamanho e prejudica a circulação do sangue no cérebro. (D) área cerebral atingida pela falta de circulação sanguínea – Saadi (2014).

A insuficiência cardíaca também se enquadra como doença crônica não transmissível, que de acordo com Schlabendorff (2014) é um conjunto de alterações clínicas que ocorrem quando o coração não consegue bombear o sangue rico em nutrientes e oxigênio para todo o corpo. Isso faz o coração trabalhar com maior intensidade para tentar superar essa deficiência e em consequência, surgem alterações clínicas como a dificuldade de respirar ou “fôlego curto”, edema (inchaço) nas pernas, pés ou tornozelos, tontura, cansaço (fadiga) e confusão mental.

Além das doenças anteriormente apresentadas, tem-se também doenças congênitas, ou seja, engloba alterações no coração e nos grandes vasos presentes no nascimento do bebê. A comunicação interventricular é o tipo mais diagnosticado. Ela se caracteriza por uma abertura na parede que divide os ventrículos (câmaras que bombeiam o sangue) do coração (Arrieta, 2017).

Completando o rol de doenças mencionadas pela OMS, trombose, que tem suas características apresentadas pelo Ministério da Saúde do Brasil: ocorre quando

há formação de um coágulo sanguíneo em uma ou mais veias grandes das pernas e das coxas. Esse coágulo bloqueia o fluxo de sangue e causa inchaço e dor na região.

O problema maior é quando um coágulo se desprende e se movimenta na corrente sanguínea, em um processo chamado de embolia. Uma embolia pode ficar presa no cérebro, nos pulmões, no coração ou em outra área, levando a lesões graves.

2.2 *Knowledge Discovery In Databases*

Para o desenvolvimento deste trabalho será utilizado o processo KDD definido por Fayyad *et. al* (1995) e, assim, sendo as etapas são explicitadas a seguir:

- **Seleção de Dados:** nesta fase, é escolhido o conjunto de dados que se pretende analisar, definindo assim os atributos e os eventos (registros).
- **Limpeza dos dados ou pré-processamento:** é a fase que determina a qualidade dos dados, onde são eliminados dados redundantes, ruídos possíveis de serem detectados e discrepância nos dados.
- **Transformação dos dados:** após o pré-processamento dos dados, estes precisam ser armazenados e formatados de forma adequada à aplicação do algoritmo na próxima fase. Também é nesta fase que são determinados atributos faltantes que podem ser obtidos de outros atributos como, por exemplo, a duração de certo evento por meio de horário inicial e horário final da ocorrência dele.
- **Data Mining:** esta é a etapa mais importante de todo o processo KDD, uma que é neste momento que se aplicam técnicas para análise dos dados por meio de algoritmos, heurísticas e metaheuristicas para a descoberta de padrões. O tempo de execução desta fase deve ser compatível na espera da solução. Muitos são os métodos, sendo que alguns dos mais conhecidos são Árvores de Decisão, Redes Neurais e Algoritmos Genéticos.

- **Interpretação do conhecimento gerado:** após a fase de DM, deve-se interpretar o conhecimento apresentado, verificando a relevância (ou não) na obtenção dos padrões e com isso, analisar a eficácia do método aplicado na etapa de DM.

Assim, pela estrutura de fases do KDD é possível percorrer caminho lógico sem considerar variáveis importantes, pautada na sequência de etapas que prezam lapidação e tratamento dos dados, com a proposta de identificar padrões. Deste modo, identifica-se os padrões na temática do presente trabalho, que possam contribuir com a qualidade de vida e bem-estar, impactando positivamente para o alcance da meta estabelecida pela OMS.

2.2.1. Tipos de Aprendizagem de Máquina

Durante o processo de Data Mining são realizadas estatisticamente algumas tarefas. As seis mais conhecidas são: descrição, classificação, estimação, predição, agrupamento e associação (LAROSE, 2005). Para o presente trabalho, será considerada apenas a tarefa de classificação para o atendimento dos objetivos deste trabalho. Segundo Araújo (2019), a classificação busca descobrir a qual classe pertence um determinado registro. Entre vários exemplos possíveis do uso da tarefa de classificação no setor público brasileiro, tem-se a descoberta do local em que uma doença pode se manifestar e a identificação de uma pessoa suspeita que possa colocar em risco a segurança dos demais em um determinado contexto da sociedade.

A seguir, serão exemplos de aprendizagem supervisionada e não supervisionada. É importante destacar a diferença entre supervisionado e não supervisionado que segundo Kubat (2017), o aprendizado supervisionado se concentra na indução de classificadores, o aprendizado não supervisionado está interessado em descobrir propriedades úteis dos dados disponíveis.

Em complemento a visão de Kubat, Igual e Seguí (2017) menciona que, aprendizagem não supervisionada é definida como a tarefa realizada por algoritmos que aprendem a partir de um conjunto de treinamento de exemplos não marcados ou

não anotados, usando os recursos das entradas para categorizar de acordo com alguns critérios geométricos ou estatísticos.

De acordo com Igual e Seguí (2017), o aprendizado de máquina envolve programas de codificação que ajustam automaticamente seu desempenho de acordo com sua exposição às informações contidas nos dados. Este aprendizado é alcançado através de um modelo com parâmetros ajustáveis que são analisados de acordo com diferentes critérios de desempenho. O aprendizado de máquina pode ser considerado um subcampo da inteligência artificial (IA) e pode-se dividir aproximadamente o campo em classes principais:

- **Aprendizagem supervisionada:** Algoritmos que aprendem a partir de conjunto de treinamento de exemplo para generalizar para o conjunto com novas instâncias (conjunto de teste). Exemplos de técnicas de aprendizagem supervisionada: regressão logística, regressão linear, máquinas de suporte para regressão ou classificação, árvores de decisão, floresta aleatória, etc.
- **Aprendizagem não supervisionada:** Algoritmos que aprendem com um conjunto de treinamento de exemplos sem um rótulo (label) de saída. Usado para explorar dados de acordo com alguns dados estatísticos, geométricos ou similaridade. Exemplos de aprendizagem não supervisionada incluem agrupamento *k-means*, agrupamento hierárquico, regras de associação, etc.

Como comentado anteriormente, as duas tarefas de aprendizagem supervisionada podem ser do tipo:

- **Regressão:** Realiza a predição de uma saída com valor numérico.
- **Classificação:** Realiza a predição de uma saída com valor categórico (nominal ou ordinal).

Assim, como na aprendizagem supervisionada há a divisão em classes, o mesmo se aplica a aprendizagem não supervisionada, que de acordo com Igual e Segui (2017). As principais tarefas de aprendizagem não supervisionadas podem ser resumidas como aquelas que abordam os seguintes grupos de problemas:

- ***Clustering*:** Tem como objetivo particionar o conjunto de exemplos em grupos.
- **Redução da dimensionalidade:** Visa reduzir a dimensionalidade dos dados.
- **Detecção de *outlier*:** Tem como objetivo encontrar eventos incomuns (por exemplo: um mau funcionamento), que distinguem parte dos dados do resto de acordo com certos critérios.

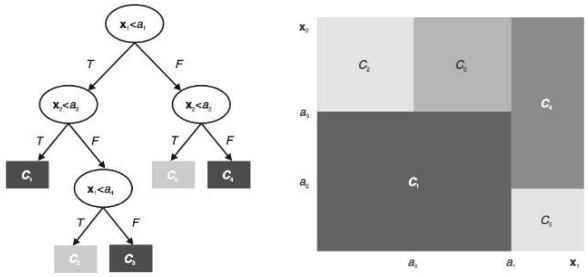
2.2.2. Árvores de Decisão

É um modelo sucessivo que une uma série de testes básicos de forma eficiente e de forma coesa, onde um atributo é testado a partir de um nó da árvore (DAMINIK *et al*.,2012).

O algoritmo árvore de decisão faz parte da aprendizagem supervisionada, e seu principal objetivo é construir um modelo de treinamento que pode ser usado para prever a classe ou valor de variáveis alvo por meio de regras de decisão de aprendizagem inferidas pelos dados de treinamento (ZHAO *et al*.,2008).

Para Faceli *et al*. (2011), a figura 3 representa uma árvore de decisão e a divisão correspondente no espaço definido pelos atributos $x1$ e $x2$. Cada nó da árvore corresponde a uma região nesse espaço. As regiões definidas pelas da árvore são mutuamente excludentes, e a reunião dessas regiões cobre todo o espaço definido pelos atributos. A interseção das regiões abrangidas por quaisquer duas folhas é vazia. A união de todas as regiões (todas as folhas) é u.

Figura 3: Uma árvore de decisão e as regiões de decisão no espaço de objetos.



Fonte: Faceli *et al*. (2011)

Ainda, para Faceli *et al*. (2011), uma árvore de decisão abrange todo o espaço de instâncias. Esse fato implica que uma árvore decisão pode fazer predições para qualquer exemplo de entrada.

2.2.3. Floresta Aleatória

São conjuntos de árvores de decisão (ADs) criadas a partir de uma base de dados. O principal desafio ao gerar uma floresta aleatória é como obter uma boa variabilidade nas árvores que a compõem, levando a um maior poder de generalização (classificar instancias desconhecidas) para o modelo.

Para se formar uma floresta, a variabilidade é obtida de duas maneiras: a) cada árvore na floresta é treinada como um subconjunto das instancias da base de dados, amostrado aleatoriamente com repetição, e b) em cada nó interno das árvores, um subconjunto dos atributos da base de dados é amostrado para que a função de divisão avalie apenas aqueles atributos (Riqueti *et al*.,2018).

Uma das vantagens em se aplicar floresta aleatória é sua eficácia em estimar os dados faltantes e manter a precisão mesmo quando grande parte dos dados estão faltando. Além disso, possibilita a utilização de grande volume de dados, podendo assim, empregar milhares de variáveis e identificar as mais significativas (as que possuem alto grau de importância). Dessa forma, a floresta aleatória é também considerada um método de redução de dimensionalidade (Sousa,2018).

2.2.4. Métricas

As métricas são direcionadoras para avaliação das tarefas de classificação, sendo que para esta tarefa existem métricas especificas. Em um problema binária por exemplo, existem quatro possíveis casos de acertos e erros entre as classes reais e as preditas pelo modelo:

- **Verdadeiros positivos (VP):** Quando o classificador prevê uma amostra como positiva e realmente é positiva.

- **Falsos positivos (FP):** Quando o classificador prevê uma amostra positiva, mas de fato, ela é negativa.
- **Negativos verdadeiros (VN):** Quando o classificador prevê uma amostra negativa e realmente é negativa.
- **Falsos negativos (FN):** Quando o classificador prevê uma amostra como negativa, mas de fato, ela é positiva.

Estas informações podem ser resumidas em uma matriz, conhecida como matriz de confusão, como descrita na Tabela 1:

Tabela 1: Matriz de Confusão

	Classe real a qual pertence	
	Classe Positiva	Classe Negativa
Predição Positiva	VP	FP
Predição Negativa	FN	VN

A combinação desses elementos permite definir várias métricas de desempenho como seguem nas equações (1) a (4):

$$\text{Acurácia:} \frac{VP + VN}{VP + VN + FP + FN} \quad (1)$$

A acurácia (equação 1), indica o quão bom é o modelo de classificação.

Em termos de colunas, encontram-se estas duas métricas parciais de desempenho:

$$\textbf{Sensibilidade:} \frac{VP}{VP + FN} \quad (2)$$

A métrica da equação 2, apresenta o quão bom é o modelo para classificar os verdadeiros positivos. Já a métrica de especificidade (equação 3) apresenta o quão bom é o modelo para classificar os verdadeiros negativos:

$$\text{Especificidade:} \frac{VP}{VP + FN} \quad (3)$$

Por fim, destaca-se a métrica de precisão:

$$\text{Precisão: } \frac{VP}{VP + FP} \quad (4)$$

A equação 4 indica quantos o modelo classificador conseguiu predizer corretamente as instâncias positivas, dentre todas as instâncias que ele classificou como positivas.

$$\text{Valor preditivo: } \frac{VN}{VN + FN} \quad (5)$$

Já a métrica de valor preditivo negativo é a que classifica o resultado negativo como de fato negativo.

2.3 Trabalhos Correlatos

Tendo em vista a magnitude dos desafios da saúde, é imprescindível a descrição e apontamentos de alguns estudos e pesquisas nessa área, estando todos organizados (na Tabela 2) alguns estudos no âmbito da saúde sob a ótica cardiológica com ênfase em acidente vascular cerebral.

Tabela 2: Trabalhos Correlatos

Autor (ano)	**Objetivo**	**Onde (base)**	**Técnicas (DM)**
Cemil Colak, Esra Karaman, M. Gokhan Turtay (2015).	Predizer o desfecho do AVC por meio dos métodos KDP, RNA e modelos SVM.	Departamento de medicina de emergência, Centro de Medicina Turgut Ozal, Universidade Inonu, Malatya, Turquia.	Redes Neurais Artificiais (RNA) *Support Vector Machine (SVM)*
Ahmet Kadir Arslan, Cemil Colak, Mehmet Ediz Sarihan (2016).	Avaliar diferentes abordagens de mineração de dados médicos para prever acidente vascular cerebral isquêmico.	Departamento de medicina de emergência, Centro de Medicina Turgut Ozal, Universidade Inonu, Malatya, Turquia.	*Support Vector Machine (SVM)* Reforço de Gradiente Estocástico (SGB) Regressão Logística Penalizada (RLP)
Wan-Yin Lin, Chun-Hsien Chen *et al* (2018).	Prever atividades de vida diária pós-AVC por meio de uma abordagem baseada	Departamento de Medicina Física e Reabilitação, Hospital Memorial Chang Gung	Regressão Logística

	em aprendizado de máquina para iniciar a reabilitação	em Linkou, Taoyuan City, Taiwan	*Support Vector Machine (SVM)* Floresta aleatória
Sang Min Sung, Yoon Jung Kang (2020).	Avaliar a aplicabilidade de algoritmos de aprendizado de máquina para prever a deterioração neurológica precoce (END) em pacientes com AVC agudo menor.	Pusan National University Hospital, Coréia do Sul	Árvores Impulsionadas Floresta de Decisão de *Bootstrap* Rede Neural Profunda Regressão Logística
Dheepitha Babu, Vandhana Karunakaran *et al* (2021).	Propor uma estratégia baseada em aprendizado de máquina para antecipar o derrame cardíaco de melhor precisão comparando a classificação administrada de cálculos de aprendizado de máquina.	Não mencionado no estudo.	*Naive Bayes* Árvores de Decisão Floresta Aleatória
Jeoung Kun Kim, Yoo Jin Choo, Min Cheol Chang (2021).	Desenvolver um modelo de rede neural profunda (DNN) e aplicar 2 algoritmos de ML bem conhecidos, regressão logística e floresta aleatória, na previsão do resultado motor 6 meses após o AVC.	Hospital Universitário Coréia do Sul	Regressão Logística Floresta Aleatória
Takeshi Imura PhD, Haruki Toda PhD, Yuji Iwamoto M.S, Tetsuji Inagawa MD *et al*. (2021).	Avaliar cinco algoritmos de aprendizado de máquina supervisionado para a classificação da possibilidade de alta domiciliar em pacientes com AVC.	Departamento de Reabilitação, Faculdade de Ciências da Saúde, Hiroshima, Japão	Árvore de Decisão *Support Vector Machine (SVM)* Análise Discriminante Linear Vizinho mais próximo Floresta aleatória

Junjie Liu, Yiyang Sun (2021).	Analisar três modelos de níveis de risco de AVC a partir de métodos de aprendizado de máquina	Censo do Hospital Popular de Shanxi Censo da comunidade de Shanxi, China	Árvore de Decisão Floresta aleatória Modelo logístico *Support Vector Machine* (SVM)
Xinyi Zhao, Xingmei Chen *et al*. (2021)	Usar diferentes algoritmos de aprendizado de máquina para identificar um modelo ideal de microRNA integrando os dados de expressão de microRNAs pré-selecionados para discriminar pacientes com acidente vascular cerebral isquêmico.	Departamento de Encefalopatia do Primeiro Hospital Afiliado da Universidade de Medicina Chinesa de Guangxi	Rede Neural Artificial Floresta aleatória *Support Vector Machine (SVM)* XGBoost
Shon Thomas, Paula de la Pena, Liam Butler *et al* (2021)	Avaliar a capacidade preditiva de vários algoritmos de aprendizado de máquina para a presença de oclusões de grandes vasos (LVO)e candidatura à trombectomia mecânica (MT) no centro de AVC abrangente.	Universidade Loyola de Medicina Stritch da Universidade de Chicago	Regressão Logística Floresta aleatória Árvore de decisão

Fonte: O autor (2021).

Em maio de 2015 foi publicado no periódico Métodos e Programas de Computador em Biomedicina, um estudo sobre aplicação do KDD, cujo objetivo foi de prever o resultado de um AVC, através da utilização de redes neurais artificiais (RNA) e modelos de máquina de vetor de suporte (*Support Vector Machines – SVM*). Os prontuários de 297 indivíduos (130 doentes e 167 saudáveis) foram adquiridos nas bases de dados do serviço de urgência e emergência do Departamento do Centro de Medicina de Turgut Ozal, Universidade Inonu, Malatya, Turquia. Os valores de precisão foram de 81,82% para RNA e 80,38% para SVM no conjunto de dados de treinamento (n=209) e 85,9% para RNA e 84,62% para SVM no conjunto de dados

teste (n=78), respectivamente. Os modelos RNA e SVM produziram valores de área sob a curva de 0,905 e 0,899 no conjunto de dados de treinamento e 0,928 e 0,91 no conjunto de dados de teste, consecutivamente.

No estudo de Ahmet Kadir *et al*., (2016) foram utilizadas três técnicas de classificação a saber: SVM, reforço por gradiente estocástico (SGB) e regressão logística penalizada (RLP). Os registros médicos de 80 pacientes e de 112 indivíduos saudáveis, com 17 preditores e uma variável-alvo, foram coletada s do serviço de urgência e emergência do Departamento de do Centro de Medicina Turgut Ozal, Universidade Inonu, Malatya, Turquia. Os valores de precisão com o intervalo de confiança (IC) de 95% foram 0,9789 (0,9470–0,9942) para SVM, 0,9737 (0,9397–0,9914) para SGB e 0,8947 (0,8421–0,9345) para RLP. Os valores de AUC com IC de 95% foram 0,9783 (0,9569–0,9997) para SVM, 0,9757 (0,9543–0,9970) para SGB e 0,8953 (0,8510–0,9396) para RLP.

Publicado no *International Journal of Medical Informatics* em março de 2018, o artigo intitulado: “Prever atividades de vida diária pós-AVC por meio de uma abordagem baseada em aprendizagem de máquina para iniciar a reabilitação” é resultado de estudo realizado em um Hospital de referência de Taiwan entre 2014 e 2016. As técnicas utilizadas foram: regressão logística SVM) e floresta aleatória. Um total de 313 indivíduos (homens: 208; mulheres: 105) foram incluídos no estudo. O desempenho dos algoritmos regressão logística e floresta aleatória foi superior (área sob a curva *Receiver Operating Characteristic* ROC (AUC): 0,79 do que o algoritmo SVM (AUC: 0,77).

No estudo de Sang Min Sung *et al*. (2020), cujo objetivo foi avaliar a aplicabilidade de algoritmos de aprendizado de máquina para prever o deterioração neurológica precoce (END) em pacientes com AVC agudo menor. Foram utilizados, quatro algoritmos de aprendizado de máquina: *boosted trees*, floresta aleatória, rede neural profunda e regressão logística. Um total de 739 pacientes foram incluídos neste estudo e os resultados mostraram que a precisão dos algoritmos *boosted trees*, floresta aleatória, rede neural profunda e regressão logística foram 0,966, 0,946, 0,966 e 0,966, respectivamente. Os valores de AUC dos algoritmos foram 0,934 (*boosted tress*), 0,932 (floresta aleatória), 0,904 (rede neural profunda) e 0,885 (regressão logística).

A proposta do estudo de Dheepitha Babu *et al.,* (2021), foi de contribuir com uma estratégia baseada em aprendizado de máquina para antecipar o derrame cardíaco. Os classificadores de aprendizado de máquina utilizados foram: *naïve bayes*, árvores de decisão e florestas aleatórias. *Naïve bayes* teve acurácia de 96,4%, enquanto a árvore de decisão de 95,9% e o classificador floresta aleatória apresentou a acurácia mais elevada, com 98,2%.

Para o estudo intitulado: “Predição da função motora em pacientes com AVC usando algoritmo de aprendizado de máquina: desenvolvimento de modelos práticos”, foi utilizado dados de 1056 pacientes, e considerado a utilização de 14 variáveis, partindo do objetivo de desenvolver um modelo de rede neural profunda e aplicação de dois algoritmos de aprendizado de máquina, regressão logística e floresta aleatória, na previsão do resultado motor seis meses após o AVC. Nesse estudo, em relação à predição da função do membro superior, para o modelo DNN, a área sob a curva (AUC) foi de 0,906. Para os modelos de regressão logística e floresta aleatória, a AUC foi de 0,874 e 0,882, respectivamente. Para a previsão da função dos membros inferiores, para os modelos: redes neurais profundas, regressão logística e floresta aleatória, as acurácias foram: 0,822, 0,768 e 0,802, respectivamente.

No estudo de Takeshi Imura *et al*. (2021), cujo objetivo foi de avaliar cinco algoritmos de aprendizado de máquina supervisionado para a classificação da possibilidade de alta domiciliar em pacientes com AVC. Os algoritmos utilizados foram: árvore de decisão, análise discriminante linear, SVM, floresta aleatória, k-vizinhos mais próximos (k-NN), sendo utilizado dados de 481 pacientes com AVC do Departamento de Reabilitação, Faculdade de Ciências da Saúde, Hiroshima, Japão. O modelo k-NN teve a melhor acurácia de classificação (84,0%) com AUC moderada (0,88) e pontuação F1 (87,8). O modelo SVM também apresentou alta precisão de classificação (82,6%) junto com a maior AUC (0,91), sensibilidade (94,4), valor preditivo negativo (87,5) e razão de verossimilhança negativa (0,088). Os modelos árvores de decisão, análise discriminante linear e floresta aleatória tiveram alta precisão de classificação ($\geq$ 79,9%) com AUCs moderadas ($\geq$ 0,84) e pontuações F1 ($\geq$ 83,8).

No estudo de Junjie Liu *et al*., (2021) foram utilizadas quatro técnicas de classificação, a saber: árvore de decisão, floresta aleatória, regressão logística e SVM. Foram utilizados dois conjuntos de dados de pesquisa de 2017 a 2020, o primeiro, censo do Hospital Popular de *Shanxi*, total de 2000 pacientes hospitalizados com AVC em 2018, e o segundo, censo na comunidade, 27.583 residentes durante o período de 2017 a 2020, sendo categorizado em baixo risco (11739), médio risco (7630) e alto risco (8214). O modelo da árvore de decisão revela que os três principais fatores são hipertensão (0,4995), sedentarismo (0,08486) e diabetes mellitus (0,07889), e o modelos de floresta aleatória mostra que os três principais fatores são: hipertensão (0,3966), hiperdipidemia (0,1229) e sedentarismo (0,1146). Para esta pesquisa, o modelo de regressão logística mostrou que as probabilidades médias de desenvolver AVC são 7.20% ± 0.55%4 para os pacientes de baixo risco, 19.02% ± 0.94% para os pacientes de nível de risco médio e 83.89% ± 0.97% para os pacientes de alto risco.

No estudo de Xinyi Zhao *et al*., (2021) a proposta foi de utilizar diferentes algoritmos de aprendizado de máquina para identificar um modelo ideal de microRNA integrando os dados de expressão de microRNAs pré-selecionados para discriminar pacientes com acidente vascular cerebral isquêmico. Os algoritmos de aprendizado de máquina utilizados foram: rede neural artificial , floresta aleatória, SVM . A acurácia de rede neural foi de 0,871, 0,948, 0,882, floresta aleatória atingiu 0,768, 0,854, 0,888 e SVM alcançou 0,910, 0,958, 0,941.

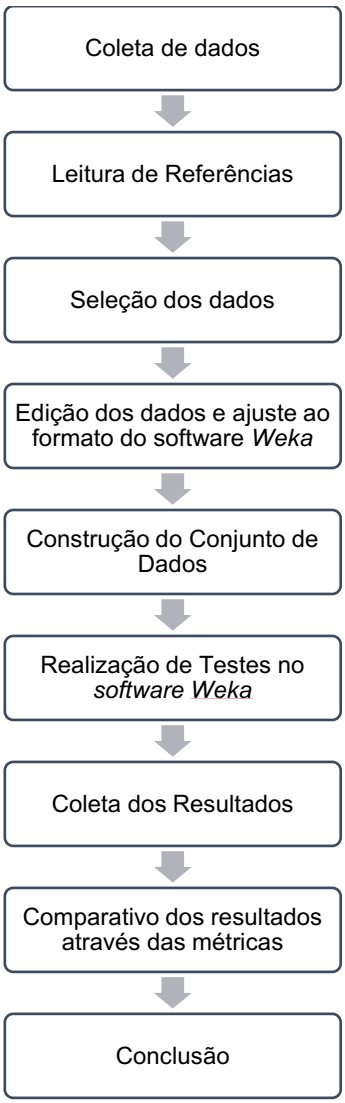
No estudo de Shon Thomas *et al*. (2021), pacientes com AVC isquêmico tratados no Loyola University Medical Center de julho de 2018 a junho de 2019 (N = 286) foram incluídos. Trinta e cinco variáveis clínicas e demográficas foram analisadas usando algoritmos de aprendizado de máquina, incluindo regressão logística, aumento de gradiente extremo, floresta aleatória, e árvores de decisão para construir modelos preditivos de presença de oclusões de grandes vasos e análise de candidatura à trombectomia mecânica. Ao usar todas as 35 variáveis, a regressão logística teve uma acurácia média melhor que a floresta aleatória.

3. METODOLOGIA

3.1 Etapas do Processo

O fluxograma (Figura 4) abaixo apresenta as etapas do processo, desde a identificação do problema, até a aplicação das técnicas através do *software Weka*, avaliação dos resultados e conclusão:

Figura 4: Etapas do processo de pesquisa



Fonte: o autor (2021)

3.2 Descrição do Conjunto de Dados

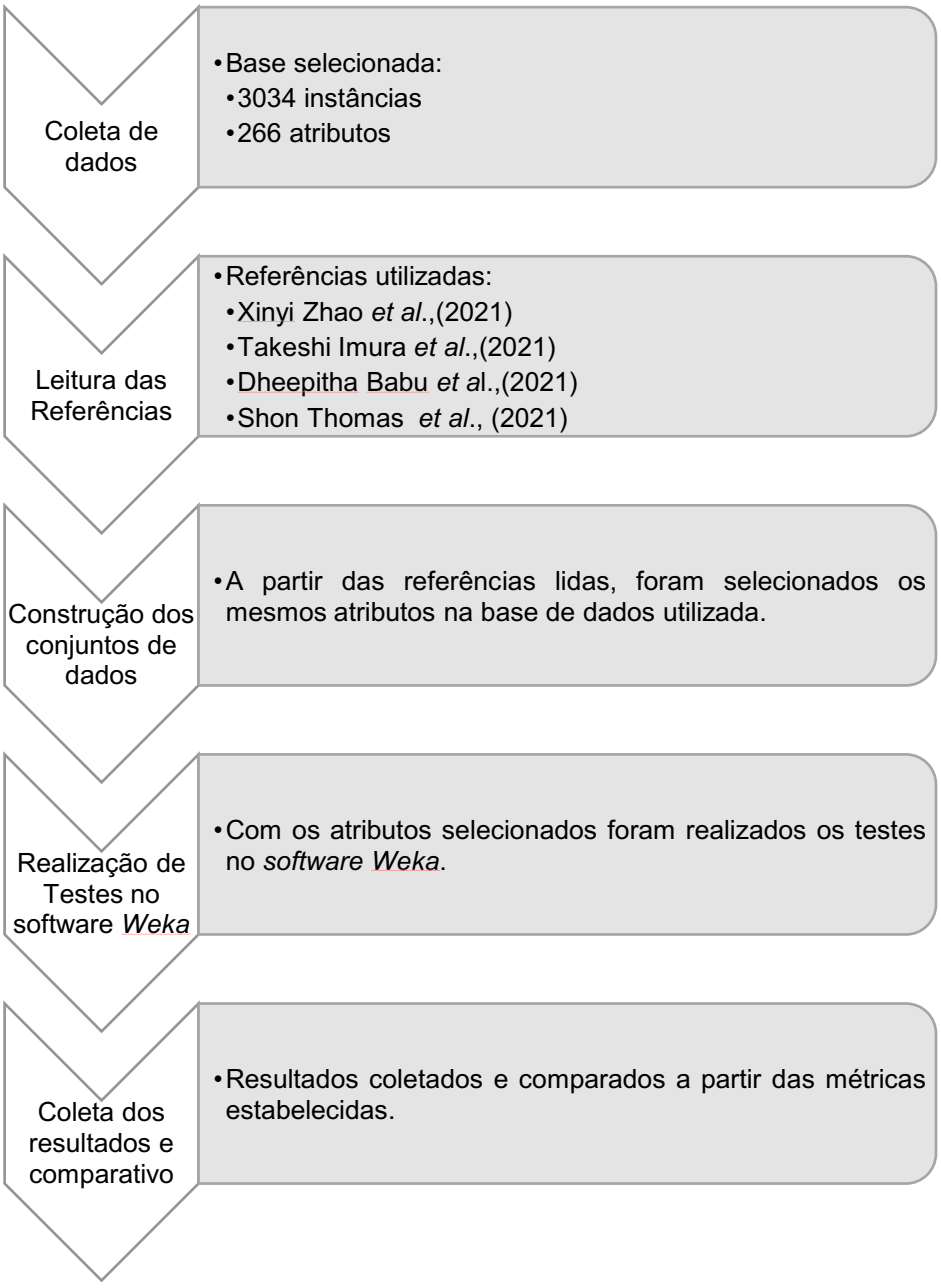
Utilizou-se a base de dados disponibilizada pelo Edinburgh *Data Share* que é um repositório digital de dados de pesquisa produzidos na Universidade de Edimburgo, Escócia, hospedado por *Information Services*. Disponível em: https://datashare.ed.ac.uk/handle/10283/1931.

Se trata de uma base de dados, intitulada “O terceiro ensaio de derrame (IST-3), cujo conjunto de dados, tem o recorte de tempo do período de 2000 a 2015, um estudo controlado randomizado em grande escala de terapia trombolítica intravenosa com o medicamento Alteplase para pacientes com AVC isquêmico agudo. O conjunto de dados inclui vários arquivos que descrevem o conjunto de dados IST-3.

3.3 Pré-Processamento

Nessa etapa foram realizados 5 passos, conforme apresentados na Figura 5:

Figura 5: Etapas do pré-processamento



Fonte: O autor (2021).

3.4 Weka

Desenvolvido na Universidade neozelandesa de Waikato, o *Waikato Environment for Knowledge Analysis* – WEKA (Disponível em: https://www.cs.waikato.ac.nz/ml/weka/), foi o software utilizado para realizar o processamento dos dados. É uma importante ferramenta, cujo objetivos são:

- Tornar as técnicas de *machine learning* geralmente disponíveis;

- Aplicá-los a problemas práticos que são importantes para a indústria da Nova Zelândia;
- Desenvolver novos algoritmos de aprendizado de máquina e fornecê-los ao mundo;
- Contribuir para um referencial teórico para o campo;

Contemplando uma coleção de algoritmos de aprendizado de máquina para tarefas de mineração de dados, ele contém ferramentas para a preparação de dados e para as tarefas de classificação, regressão, agrupamento, mineração de regras de associação e visualização.

Assim, por meio dessa coleção disponível foi possível processar os dados utilizados, aplicar as técnicas disponíveis e verificar os modelos e suas respectivas métricas de avaliação.

4. RESULTADOS E DISCUSSÃO

Neste capítulo são discutidos os resultados do trabalho, bem como as discussões em cima de cada teste realizado.

4.1 Preparação do conjunto de dados

A partir dos trabalhos correlatos anteriormente citados, quatro deles foram selecionados como direcionadores da seleção de atributos para o processamento da base de dados disponibilizada pelo *Edinburgh Data Share*. São eles (Tabela 3).

Tabela 3: Trabalhos correlatos direcionadores

Titulo	Autor (ano)	Quantidade de atributos
Análise de aprendizado de máquina de dados de expressão de Micro RNA revela novo bio marcador de diagnóstico para AVC isquêmico	Xinyi Zhao, Xingmei Chen *et al*.,(2021)	5
Comparação de algoritmos de aprendizado de máquina supervisionados para classificação da possibilidade de alta domiciliar em pacientes com AVC convalescente: uma análise secundária	Takeshi Imura, Haruki Toda *et al*. ,(2021	5
Previsão baseada em Gui de derrame cardíaco usando inteligência artificial	Dheepitha Babu, Vandhana Karunakaran *et al*., (2021)	8
Avaliar a capacidade preditiva de vários algoritmos de aprendizado de máquina para a presença de oclusões de grandes vasos (LVO)e candidatura à trombectomia mecânica (MT) no centro de AVC abrangente.	Shon Thomas *et al*., (2021)	15

Fonte: O autor (2021).

4.2 Comparação dos Resultados

Os resultados dos modelos estão apresentados nas Tabelas 4.

Tabela 4: Atributos de Xinyi Zhao *et al*., (2021)

Métricas	Árvore de Decisão	Floresta Aleatória
Acurácia	68,12 %	69,08 %
Precisão	0,749	0,751
Especificidade	0,547	0,540
Sensibilidade	0,756	0,775

Fonte: O autor (2021).

Considerando os atributos que tiveram como referência o estudo de Xinyi Zhao *et al*., (2021) o modelo de floresta aleatória teve melhor performance em relação a árvore de decisão nas métricas de acurácia e sensibilidade. Ou seja, o quão bom o modelo de floresta aleatória foi assertivo para indicar a evolução para óbito, bem como o quão bom o modelo foi assertivo para indicar os verdadeiros positivos. Por outro lado, a identificação dos verdadeiros negativos (especificidade) teve maior assertividade no modelo de árvore de decisão. Com relação a matriz de confusão obtida, apresenta o desempenho para o modelo de árvore de decisão e floresta aleatória, respectivamente:

Tabela 5: Matriz de Confusão dos atributos de Xinyi Zhao *et al*., (2021) para árvore de decisão

	Classe real a qual pertence	
	Classe Positiva	Classe Negativa
Predição Positiva	1472	474
Predição Negativa	493	595

Fonte: O autor (2021).

Os resultados obtidos, indicam que o modelo conseguiu predizer de verdadeiro positivo, total de 1472, enquanto falso negativo foi de 474. Ou seja, conseguiu ter uma boa performance considerando os atributos do referido referencial para o modelo de árvore de decisão.

Tabela 6: Matriz de Confusão dos atributos de Xinyi Zhao *et al*., (2021) para floresta aleatória

	Classe real a qual pertence	
	Classe Positiva	Classe Negativa
Predição Positiva	1509	437
Predição Negativa	501	587

Fonte: O autor (2021).

Por outro lado, alterando de árvore de decisão para floresta aleatória e considerando os mesmos atributos, percebe-se que o modelo conseguiu predizer de verdadeiro positivo, total de 1509, ou seja, melhor performance que árvore de decisão, enquanto falso negativo foi de 474 para 437 (Tabela 6).

Na tabela 7, foram considerados os atributos do estudo de Dheepitha Babu et.,al (2021). Os resultados obtidos foram:

Tabela 7: Atributos de Dheepitha Babu *et al*., (2021)

Métricas	Árvore de Decisão	Floresta Aleatória
Acurácia	69,67 %	68,49 %
Precisão	0,749	0,741
Especificidade	0,523	0,510
Sensibilidade	0,794	0,783

Fonte: O autor (2021).

Analisando o desempenho, do conjunto 2 (Tabela 7), nota-se que de acordo com as métricas estabelecidas, para ambos os modelos os resultados foram próximos. Vale ressaltar que o número de verdadeiro positivo, chegou em 1545, para árvore de decisão e 1523 para floresta aleatória, conforme apresentado nas tabelas (8 e 9).

Tabela 8: Matriz de Confusão dos Atributos de Dheepitha Babu *et al*., (2021) para árvore de decisão

	Classe real a qual pertence	
	Classe Positiva	Classe Negativa
Predição Positiva	1545	401
Predição Negativa	519	569

Fonte: O autor (2021).

Tabela 9: Matriz de Confusão dos Atributos de Dheepitha Babu *et al*., (2021) para floresta aleatória

	Classe real a qual pertence	
	Classe Positiva	Classe Negativa
Predição Positiva	1523	423
Predição Negativa	523	555

Fonte: O autor (2021).

A partir dos resultados obtidos na matriz de confusão de ambos os modelos, percebe-se que a assertividade dos verdadeiros positivos é próxima.

Tabela 10: Atributos de Takeshi Imura *et al*., (2021)

Métricas	Árvore de Decisão	Floresta Aleatória
Acurácia	69,14 %	69,34 %
Precisão	0,767	0,766
Especificidade	0,596	0,588
Sensibilidade	0,745	0,752

Fonte: O autor (2021).

Na sequência, (Tabelas 11 e 12), tem-se as matrizes de confusão de confusão, onde são apresentados o desempenho para as técnicas utilizadas.

Tabela 11: Matriz de Confusão dos Atributos de Takeshi Imura *et al*., (2021) para árvore de decisão

	Classe real a qual pertence	
	Classe Positiva	Classe Negativa
Predição Positiva	1450	496
Predição Negativa	440	648

Fonte: O autor (2021).

Tabela 12: Matriz de Confusão dos Atributos de Takeshi Imura *et al*., (2021) para floresta aleatória

	Classe real a qual pertence	
	Classe Positiva	Classe Negativa
Predição Positiva	1464	482
Predição Negativa	448	640

Fonte: O autor (2021).

Considerando os resultados obtidos, percebe-se que a assertividade do modelo para verdadeiro positivo, foi na classificação de floresta aleatória, correspondendo 1464 contra 1450. Para a tabela 13, os resultados do quarto e último estudo realizado, utilizados como alicerce para esse estudo:

Tabela 13: Atributos de Shon Thomas *et al*., (2021)

Métricas	Árvore de Decisão	Floresta Aleatória
Acurácia	67,50 %	63,28 %
Precisão	0,755	0,678
Especificidade	0,544	0,309
Sensibilidade	0,730	0,814

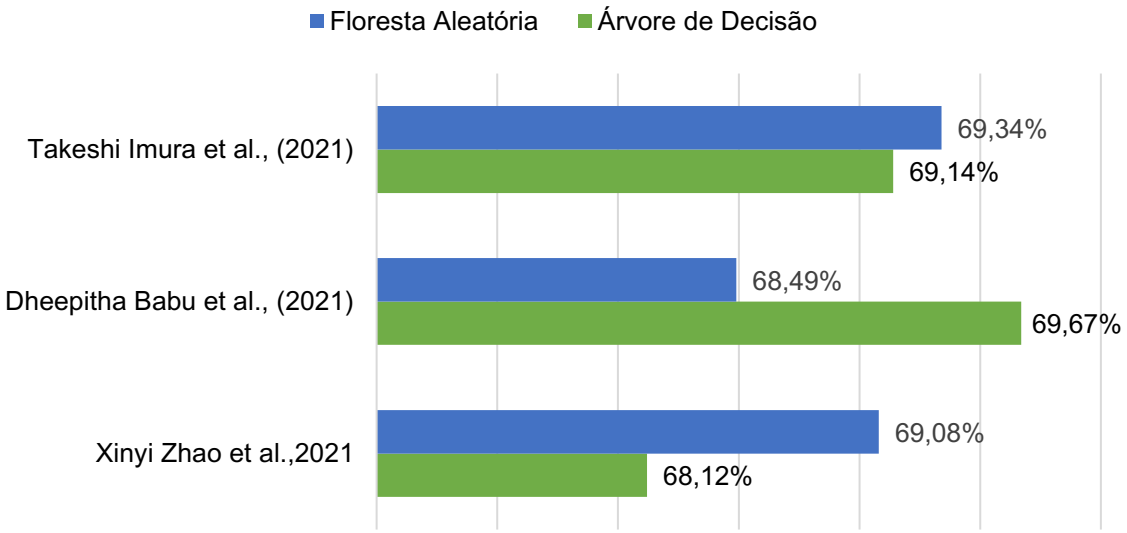
Fonte: O autor (2021).

Os estudos denominados como trabalhos correlatos direcionadores, assumiram papel fundamental para esse trabalho. Contudo, dos estudos selecionados o que não teve a melhor performance foi o que os atributos utilizados para classificação, foi de: Shon Thomas *et al*., (2021), ou seja, não é um modelo aplicável em situações reais.

4.3 Discussão

Quatro estudos foram utilizados como referência para seleção dos atributos. Exceto os resultados obtidos que tiveram como referência os atributos do estudo de Shon Thomas et al., (2021), os demais tiveram resultados similares, conforme apresentado na figura 6:

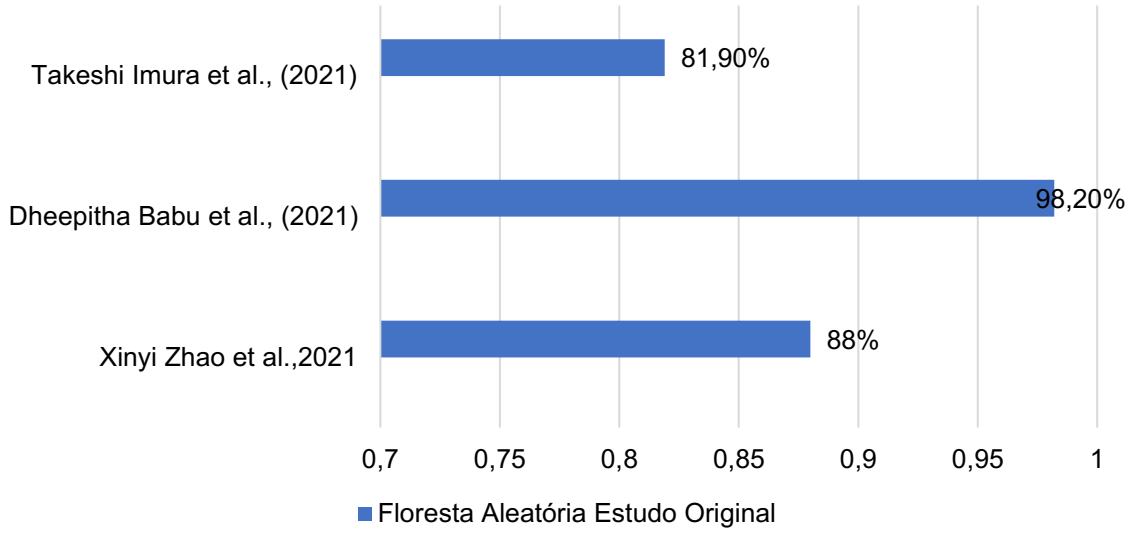
Figura 6 – Acurácia a partir dos atributos dos trabalhos correlatos direcionadores



Fonte: O autor (2021).

Por outro lado, a acurácia dos referidos estudos foi maior em relação à da base utilizada, conforme apresentado na figura 7:

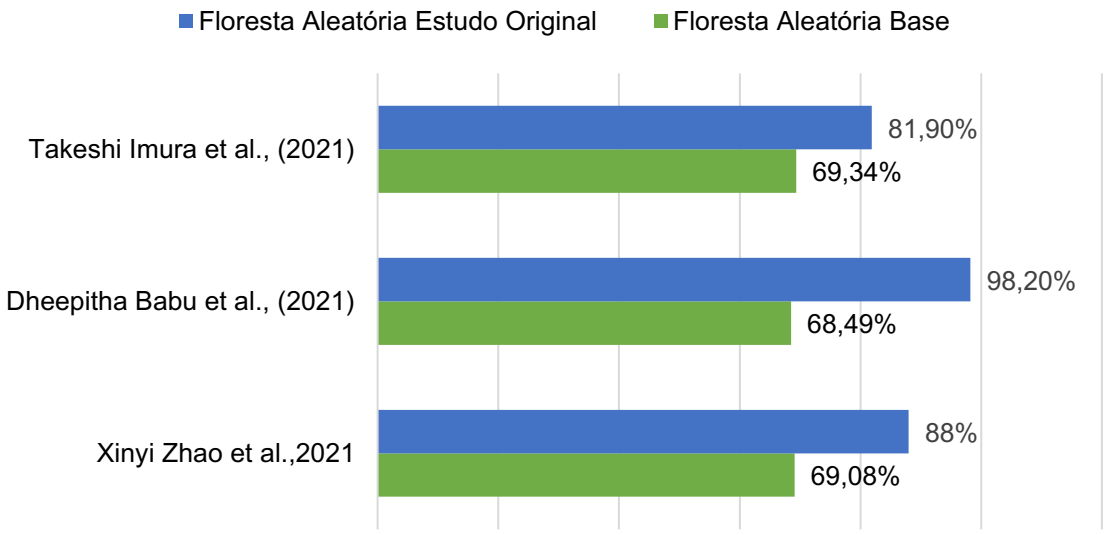
Figura 7 – Acurácia de Floresta Aleatória do Estudo Original



Fonte: O autor (2021).

É importante destacar que os referidos estudos utilizaram outras técnicas de classificação, porém, comum ao presente trabalho, é a técnica de floresta aleatória. Na sequência, figura 8 apresenta o comparativo do desempenho de floresta aleatória a partir da base utilizada no presente estudo em relação aos estudos originais.

Figura 8 – Comparativo de Acurácia de Floresta Aleatória do Estudo Original em relação a base utilizada



Fonte: O autor (2021).

Tendo em vista o comparativo apresentado na figura 8, é perceptível a diferença da acurácia, de modo que nos estudos originais, a acurácia foi superior a 80% como é o caso do estudo de Takeshi Imura *et al*. (2021), superior a 85% como apresentado no estudo de Xinyi Zhao *et al*.(2021) e superior a 95% como é o estudo de Dheepitha Babu *et al*.(2021) que teve acurácia de 98,20%.

Ou seja, comum aos três estudos, a classificação do modelo tiveram resultados superiores aos encontrados nesse estudo. É importante salientar que a dimensão da base, outras técnicas e parâmetros refletem diretamente nesse desempenho. Assim sendo, pode-se utilizar a mesma base desse estudo, porém, aplicando outras técnicas e atributos.

5. CONCLUSÃO

Inúmeros são os desafios da área da saúde, por outro lado, no rol da engenharia encontram-se diferentes saberes que podem ser aplicados em distintas áreas, impactando positivamente na resolução de *gaps* encontrados quer seja na saúde ou em quaisquer outras áreas.

A proposta desse trabalho foi de olhar para um dos recortes e desafios da saúde, o AVC, inclusive estabelecido como meta pela OMS, no que tange a redução de óbitos decorrente de sua ocorrência.

Para tanto, foi pesquisado e encontrado a banco de dados internacional de testes de AVC, base que totaliza 3034 instâncias e 266 atributos. Foi feito seleção de atributos a partir de referenciais, denominados trabalhos correlatos direcionadores.

Árvore de decisão e floresta aleatória foram as duas técnicas de classificação utilizadas, de modo que, o melhor resultado foi o que teve como referência os atributos utilizados no estudo de Dheepitha Babu *et.,al*., (2021), os resultados obtidos foram: 69,67 de acurácia, precisão de 0,749, seguido pela sensibilidade de 0,794.

Pela grandeza e dimensão da base, recomenda-se como trabalhos futuros, a utilização de outros atributos da mesma base de dados, de modo a investigar novas possibilidades das técnicas de classificação utilizadas no presente estudo, bem como, a utilização de outros parâmetros.

REFERÊNCIAS

ABDAR, Moloud; KSIąŜEK, Wojciech; ACHARYA, U Rajendra; TAN, Ru-San; MAKARENKOV, Vladimir; PŁAWIAK, Paweł. A new machine learning technique for an accurate diagnosis of coronary artery disease. **Computer Methods And Programs In Biomedicine**, [S.L.], v. 179, p. 104992, out. 2019. Elsevier BV. http://dx.doi.org/10.1016/j.cmpb.2019.104992. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0169260718314585#bib0050. Acesso em: 12 abr. 2020.

ARSLAN, Ahmet Kadir; COLAK, Cemil; SARIHAN, Mehmet Ediz. Different medical data mining approaches based prediction of ischemic stroke. **Computer Methods And Programs In Biomedicine**, [S.L.], v. 130, p. 87-92, jul. 2016. Elsevier BV. http://dx.doi.org/10.1016/j.cmpb.2016.03.022. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0169260716301067. Acesso em: 31 out. 2021.

BABU, Dheepitha; KARUNAKARAN, Vandhana; GOPINATH, Swathi; KIRUBHA, S.P. Angeline; LATHA, S.; MUTHU, P.. Gui based prediction of heart stroke using artificial intelligence. **Materials Today**: Proceedings, [S.L.], v. 47, p. 104-108, 2021. Elsevier BV. http://dx.doi.org/10.1016/j.matpr.2021.03.693. Disponível em: https://www.sciencedirect.com/science/article/pii/S2214785321027954. Acesso em: 31 out. 2021.

Brasil. Ministério da Saúde (MS). **Plano de ações estratégicas para o enfrentamento das doenças crônicas não transmissíveis (DCNT) no Brasil 2011-2022** [Internet]. Brasília: MS; 2011. [acesso em 2021 Set 27]. Disponível em: http://bvsms.saude.gov.br/bvs/publicacoes/plano_acoes_enfrent_dcnt_2011.pdf

Barros, N. *et al*. **Entendendo as doenças cardiovasculares**. Porto Alegre: Artmed, 2014.

COLAK, Cemil; KARAMAN, Esra; TURTAY, M. Gokhan. Application of knowledge discovery process on the prediction of stroke. **Computer Methods And Programs In Biomedicine**, [S.L.], v. 119, n. 3, p. 181-185, maio 2015. Elsevier BV. http://dx.doi.org/10.1016/j.cmpb.2015.03.002. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0169260715000565. Acesso em: 31 out. 2021.

FACELI. Katti. *et al*. **Inteligência Artificial: Uma Abordagem de Aprendizagem de**

Máquina. Rio de Janeiro: LTC, 2011.

GOES.Anderson Roges Teixeira. STEINER. Maria Teresinha Arns. **O Processo KDD Aplicado na Extração de Regras: Um Estudo de Caso da Área Médica.** Congresso Latino-Iberoamericano de Investagácion Operativa. Simpósio Brasileiro de Pesquisa Operacional. Rio de Janeiro, RJ, Brasil, setembro de 2012.

IGUAL.L.SEGUÍ.S. **Introduction to Data Science A Python Approach to Concepts, Techniques and Applications**. Springer International Publishing Switzerland 2017.

IMURA, Takeshi; TODA, Haruki; IWAMOTO, Yuji; INAGAWA, Tetsuji; IMADA, Naoki; TANAKA, Ryo; INOUE, Yu; ARAKI, Hayato; ARAKI, Osamu. Comparison of Supervised Machine Learning Algorithms for Classifying of Home Discharge Possibility in Convalescent Stroke Patients: a secondary analysis. **Journal Of Stroke And Cerebrovascular Diseases**, [S.L.], v. 30, n. 10, p. 106011, out. 2021. Elsevier BV. http://dx.doi.org/10.1016/j.jstrokecerebrovasdis.2021.106011. Disponível em: https://www.sciencedirect.com/science/article/pii/S105230572100416X. Acesso em: 31 out. 2021.

KIM, Jeoung Kun; CHOO, Yoo Jin; CHANG, Min Cheol. Prediction of Motor Function in Stroke Patients Using Machine Learning Algorithm: development of practical models. **Journal Of Stroke And Cerebrovascular Diseases**, [S.L.], v. 30, n. 8, p. 105856, ago. 2021. Elsevier BV. http://dx.doi.org/10.1016/j.jstrokecerebrovasdis.2021.105856. Disponível em: https://www.sciencedirect.com/science/article/pii/S1052305721002597. Acesso em: 31 out. 2021.

KUBAT.M. **An Introduction to Machine Learning**. Segunda Edição. Springer International Publishing Switzerland 2017.

LONGARAY, André A. **Introdução à Pesquisa Operacional. Disponível em: Minha Biblioteca**, Editora Saraiva, 2013. Acesso em 31 out.2021.

LIN, Wan-Yin; CHEN, Chun-Hsien; TSENG, Yi-Ju; TSAI, Yu-Ting; CHANG, Ching-Yu; WANG, Hsin-Yao; CHEN, Chih-Kuang. Predicting post-stroke activities of daily living through a machine learning-based approach on initiating rehabilitation. **International Journal Of Medical Informatics**, [S.L.], v. 111, p. 159-164, mar. 2018. Elsevier BV. http://dx.doi.org/10.1016/j.ijmedinf.2018.01.002. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S1386505618300029. Acesso em: 31 out. 2021.

LIU, Junjie; SUN, Yiyang; MA, Jing; TU, Jiachen; DENG, Yuhui; HE, Ping; LI, Rongshan; HU, Fengyun; HUANG, Huaxiong; ZHOU, Xiaoshuang. Analysis of main risk factors causing stroke in Shanxi Province based on machine learning models. **Informatics In Medicine Unlocked**, [S.L.], v. 26, p. 100712, 2021. Elsevier BV. http://dx.doi.org/10.1016/j.imu.2021.100712. Disponível em: https://www.sciencedirect.com/science/article/pii/S2352914821001933#!. Acesso em: 31 out. 2021.

MITCHELL, Tom. **Machine Learning**. McGraw-Hill, 1997.

OPAS/OMS Brasil – **Organização Mundial da Saúde divulga novas estatísticas.**

Disponívelem:https://www.paho.org/pt/search/r?keys=organizacao+mundial+da+sau

de+divulga+novas+estatisticas+mundiais+de+saude+Brasil.Acesso em: 23 ago.2018.

RAGASDALE. Cliff. T. **Modelagem de planilha e análise de decisão: uma introdução prática a business analytics.** São Paulo: Cengage Learning, 2017.

RAMOS, S. *et al*. **Entendendo as doenças cardiovasculares**. Porto Alegre: Artmed, 2014. 104 p.

Sandercock, P; Wardlaw, J; Lindley, R; Cohen, G; Whiteley, W. (2016). O terceiro ensaio internacional de acidente vascular cerebral (IST-3), 2000-2015 [conjunto de dados]. Universidade de Edimburgo e Unidade de Ensaios Clínicos de Edimburgo. https://doi.org/10.7488/ds/1350. Data Share Edinburgh. Disponível em: https://datashare.ed.ac.uk/handle/10283/1931. Acesso em: Acesso em: 12 abr. 2021.

SUNG, Sang Min; KANG, Yoon Jung; CHO, Han Jin; KIM, Nae Ri; LEE, Suk Min; CHOI, Byung Kwan; CHO, Giphil. Prediction of early neurological deterioration in acute minor ischemic stroke by machine learning algorithms. **Clinical Neurology And Neurosurgery**, [S.L.], v. 195, p. 105892, ago. 2020. Elsevier BV. http://dx.doi.org/10.1016/j.clineuro.2020.105892. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0303846720302353. Acesso em: 31 out. 2021.

THOMAS, Shon; LAPENA, Paula de; BUTLER, Liam; AKBILGIC, Oguz; HEIFERMAN, Daniel M.; GARG, Ravi; GILL, Rick; SERRONE, Joseph C.. Machine learning models improve prediction of large vessel occlusion and mechanical thrombectomy candidacy in acute ischemic stroke. **Journal Of Clinical Neuroscience**, [S.L.], v. 91, p. 383-390, set. 2021. Elsevier BV. http://dx.doi.org/10.1016/j.jocn.2021.07.021. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0967586821003738. Acesso em: 31 out. 2021.

UNIVERSIDADE DE WAIKATO. **Weka 3: Machine Learning Software in Java. [S. l.], 2021**. Disponível em: https://www.cs.waikato.ac.nz/ml/weka/. Acesso em: 11 nov. 2021.

ZHAO, Xinyi; CHEN, Xingmei; WU, Xulong; ZHU, Lulu; LONG, Jianxiong; SU, Li; GU, Lian. Machine Learning Analysis of MicroRNA Expression Data Reveals Novel Diagnostic Biomarker for Ischemic Stroke. **Journal Of Stroke And Cerebrovascular Diseases**, [S.L.], v. 30, n. 8, p. 105825, ago. 2021. Elsevier BV. http://dx.doi.org/10.1016/j.jstrokecerebrovasdis.2021.105825. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S1052305721002287. Acesso em: 31 out. 2021.