UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"
Câmpus de São José do Rio Preto

IBILCE

JARDEL BRANDÃO DOS SANTOS

Predição de Dados na Saúde – Uma Abordagem de Viabilidade do modelo Zero-Shot-Learning

São José do Rio Preto
2024

JARDEL BRANDÃO DOS SANTOS

Predição de Dados na Saúde – Uma Abordagem de Viabilidade do modelo Zero-Shot-Learning

Trabalho de Conclusão de Curso apresentado ao Departamento de Ciências de Computação e Estatística do Instituto de Biociências Letras e Ciências Exatas da Universidade Estadual Paulista "Júlio de Mesquita Filho", como parte dos requisitos necessários para obtenção do grau de Bacharel em Ciência da Computação.

Prof. Dr. Carlos Roberto Valêncio

São José do Rio Preto
2024

S237p

Santos, Jardel Brandão dos

Predição de dados na saúde - uma abordagem de viabilidade do modelo Zero-Shot-Learning / Jardel Brandão dos Santos. -- São José do Rio Preto, 2024

55 p. : il., tabs.

Trabalho de conclusão de curso (Bacharelado - Ciência da Computação) - Universidade Estadual Paulista (UNESP), Instituto de Biociências Letras e Ciências Exatas, São José do Rio Preto

Orientador: Carlos Roberto Valêncio

1. Predição de dados. 2. Inteligência artificial. 3. Saúde. 4. Zero-Shot-Learning. I. Título.

Sistema de geração automática de fichas catalográficas da Unesp. Dados fornecidos pelo autor(a).

JARDEL BRANDÃO DOS SANTOS

Predição de Dados na Saúde – Uma Abordagem de Viabilidade do modelo Zero-Shot-Learning

Trabalho de Conclusão de Curso apresentado ao Departamento de Ciências de Computação e Estatística do Instituto de Biociências Letras e Ciências Exatas da Universidade Estadual Paulista "Júlio de Mesquita Filho", como parte dos requisitos necessários para obtenção do grau de Bacharel em Ciência da Computação.

Data da aprovação: -- de agosto de 2024

Banca examinadora:

Prof. Dr. Carlos Roberto Valêncio
UNESP – São José do Rio Preto
Orientador

Prof. Dr. Wallace Correa de Oliveira Casaca
UNESP – São José do Rio Preto

Prof. Dr. Geraldo Francisco Donegá Zafalon
UNESP – São José do Rio Preto

São José do Rio Preto
2024

Agradecimentos

Agradeço primeiramente a Deus e a minha família, minha mãe Lucia Galdino Brandão, meu irmão Douglas Brandão dos Santos e minha cunhada Beatriz Lima por me apoiarem e me ajudarem durante todo o curso de bacharelado.

Agradeço também ao meu falecido pai que sempre lutou para que meus estudos fossem possíveis. O mérito dessa conquista é dele.

Agradeço também aos meus amigos da faculdade de anos anteriores que tanto me auxiliaram me ajudando em estudos e me acalmando em horas de desespero acadêmico.

Por fim agradeço a mim mesmo por nunca desistir e apesar das dificuldades e tropeços do caminho, sempre me levantar e seguir em frente.

“Quando olhei para o pico da montanha, na minha cabeça, eu já havia caído. E assim, eu subi.”

Pantheon – League of Legends

Resumo

O início do século XXI tem se apresentado como período importante da história no qual ocorreram mudanças permanentes e expressivas no campo da tecnologia. O grande volume de dados gerados diariamente, advento denominado Big Data, proporciona informações e conhecimento que, quando bem-organizados e analisados, oferecem uma compreensão mais ampla e acurada sobre o que precisa ser representado. A recente ascensão da inteligência artificial tem-se provado uma importante ferramenta em diversos campos de pesquisa, dentre eles, a automação de processos de *Data Mining*, por conta do volume e dimensionalidade dos dados. Com o propósito de contribuir neste sentido, este trabalho explora uma abordagem de aprendizado de máquina, compreendido no campo da inteligência artificial, denominada *Zero-Shot Learning*, a qual foi originada para a classificação de imagens a partir de informações contextuais, para a classificação de dados tabulares da saúde onde o formato dos dados e as condições enfrentadas favorecem esse tipo de técnica. Por fim, a comparação de técnicas comumente utilizadas para classificação, com o intuito de validar a viabilidade do *Zero-Shot-Learning* neste meio.

Palavras-chave: Inteligência Artificial, Mineração de Dados, Extração de Conhecimento, Saúde, Predição de Dados, Big Data, Aprendizado de Máquina, *Zero-Shot Learning*.

Abstract

The beginning of the 21st century has proven to be a pivotal period in history, marked by permanent and significant changes in the field of technology. The large volume of data generated daily, a phenomenon known as Big Data, provides information and knowledge that, when well-organized and analyzed, offers a broader and more accurate understanding of what needs to be represented. The recent rise of artificial intelligence has proven to be an important tool in various fields of research, including the automation of Data Mining processes, due to the volume and dimensionality of the data. With the aim of contributing in this context, this work explores a machine learning approach, within the field of artificial intelligence, known as Zero-Shot Learning. Originally developed for image classification based on contextual information, this technique is applied here to classify tabular health data, where the data format and conditions encountered favor such an approach. Finally, commonly used classification techniques are compared to validate the feasibility of Zero-Shot Learning in this domain.

Keywords: Artificial Intelligence, Data Mining, Knowledge Discovery in Databases, Healthcare, Data Prediction, Big Data, Machine Learning, Zero-Shot Learning.

Lista de Ilustrações

Lista de Tabelas

Lista de Abreviaturas e Siglas

IA – *Inteligência Artificial*

ML – *Machine Learning*

KDD – *Knowledge Discovery in Databases*

DM – *Data Mining*

ZSL – *Zero-Shot Learning*

FSL – *Few-Shot Learning*

GBD – *Grupo de Banco de Dados*

IBILCE – *Instituto de Biociências, Letras e Ciências Exatas*

GPT – *Generative Pre-trained Transformer*

ICD-10-CM – *International Classification Disease, décima revisão, Modificação Clínica*

HVE – *Hipertofia Ventricular Esquerda*

CardiacSiamese-CAMZ – *CardiacSiamese- Cardiac Apical Multiview Zero-Shot*

MSK - *Memorial Sloan Kettering Cancer Center*

NLP – *Natural Language Processing*

ID – *Identificação*

Sumário

1 Introdução

Na era do crescimento significativo de informações, a velocidade de geração de informações está aumentando dia após dia, e a informação produzida pelos diversos segmentos da sociedade, somado aos novos e variados meios de comunicação, gera informações em volumes importantes (YANG et al., 2020). Ainda, com o advento da internet, uma avalanche de dados começou a ser criada diariamente, em que se proporcionam informações abrangentes e de todos os níveis de relevância. Nos últimos anos, o termo *Big Data* se tornou comum no vocabulário dos setores industrial, financeiro e da saúde (YANG et al., 2020; TRIFIRÒ; SULTANA; BATE, 2017).

Existem diversos desafios atrelados ao Big Data, um dos mais expressivos é o gerenciamento da quantidade significativa de dados que abrange. O termo *Knowledge Discovery in Databases* — KDD, "Descoberta de Conhecimento em Bancos de Dados" — refere-se a um processo não trivial para encontrar informações existentes em dados (KAMM; JAZDI; WEYRICH, 2021). Para que o processo de KDD ocorra de forma coerente e eficiente, como dito anteriormente, faz-se necessário a limpeza desses dados com o processo denominado de *Data Cleaning.* Assegurar a qualidade dos dados é, principalmente, aplicar o processo de *Data Cleaning*, o qual objetiva remover anomalias que possam afetar os dados (LATTAR; BEM; HAJJAMI, 2020).

Dessa forma, uma das ferramentas amplamente aplicadas atualmente é a intitulada Inteligência Artificial (AI), proposta por John McCarthy em 1956 para descrever máquinas que conseguem realizar funções cognitivas, bem como o aprendizado e solução de problemas que se assemelham a mente humana (GUPTA et al., 2023). A finalidade final de uma IA é aprender e identificar associações entre dados e resultados de interesse (KAMM; JAZDI; WEYRICH, 2021). Portanto, tendo em vista a quantidade de dados que precisam ser pré-processados para

que informações relevantes sejam extraídas, a IA pode ser entendida como um método para auxiliar a automatização desses processos que seriam muito mais custosos física e computacionalmente se realizados de outras formas.

1.1 Motivação e Escopo

A inteligência artificial é extremamente versátil e possui o potencial de revolucionar diversas áreas. Este trabalho, no entanto, foca especificamente na área da saúde, pois segundo Olawade et al. (2023) a IA possui potencial para transformar o domínio da área da saúde, uma vez que já apresentou resultados quando aplicada para diagnósticos de doenças, de modo a prever o desenvolvimento de doenças infecciosas e encontrar novos alvos para medicamentos.

De acordo com Shamshirband et al. (2021) durante a década passada, diversas técnicas de IA e *Machine Learning* (MLI) foram utilizadas para analisar quantidades massivas de dados da saúde. Desse modo, surge o termo Saúde Digital, o qual é utilizado para descrever o espaço multidisciplinar e inovador que abrange a medicina e a tecnologia (TAN et al., 2023). Também, a Organização Mundial da Saúde recomenda a avaliação das ferramentas de saúde digital com base nos benefícios, danos, aceitabilidade, viabilidade, uso de recursos e considerações de equidade (ORGANIZAÇÃO MUNDIAL DA SAÚDE, 2019).

Como a modelagem preditiva tem o poder de aumentar a capacidade humana de prever a propagação de doenças infecciosas e orientar tratamentos de saúde pública, é uma aplicação fundamental da inteligência artificial para a mesma (OLAWADE et al., 2023; SINGH et al., 2020). Esse fato indica a relevância da investigação de maneiras alternativas com o uso de inteligências artificiais com enfoque em classificar de forma a se alcançar melhor eficiência computacional e resultados similares ou melhores ao de mecanismos convencionais.

Os recursos de ML podem auxiliar a realização de tarefas rotineiras e sistemáticas com alta consistência, o que permite que profissionais da saúde possam dedicar mais tempo a resolver problemas clínicos mais complexos ou que exijam uma interação humana substancial (NGIAM; KHOR, 2019).

A utilização dos recursos de IA no âmbito da saúde podem proporcionar benefícios, uma vez que pode auxiliar no aumento da precisão de diagnósticos, prevenção de desenvolvimento de doenças, dentre outros fatores. Além disso, a constante pesquisa e evolução

do uso da computação juntamente ao ramo da saúde é de interesse da Organização Mundial da Saúde (ORGANIZAÇÃO MUNDIAL DA SAÚDE, 2019).

1.2 Objetivo

A abordagem de aprendizado de máquina, denominada de *Zero-Shot Learning* – ZSL, tem como objetivo treinar um modelo que possa classificar objetos de classes não vistas, pertencentes ao domínio alvo, por meio do conhecimento obtido de outras classes vistas, denominado de domínio fonte, com a ajuda de informações semânticas (POURPANAH et al., 2023). Apesar de destinado a classificação de imagens, o ZSL pode ser usado em outros contextos, dessa forma, o objetivo deste trabalho é apresentar os resultados da sua aplicação à predição sobre dados tabulares.

De modo a considerar o valor a ser extraído do uso de IA's no contexto da saúde, a utilização do ZSL pode proporcionar uma diminuição de custos computacionais e de tempo se comparado a métodos tradicionais, porém com a mesma eficiência, uma vez que o método propõe a classificação de dados sem conhecimento prévio sobre eles, alicerçado no conhecimento semântico. Portanto, o objetivo do presente trabalho é apresentar os resultados alcançados na utilização do método ZSL para realizar predições com dados da área da saúde e a respectiva comparação com os métodos de classificação convencionalmente utilizados.

1.3 Metodologia

A metodologia seguida para o desenvolvimento deste trabalho está dividida nas seguintes etapas: levantamento bibliográfico, determinação de bases de dados a serem utilizadas para classificação, aplicação dos mecanismos sugeridos e posterior análise e comparação dos resultados.

O levantamento teórico e bibliográfico realizado abordou os seguintes temas: *Big Data*, KDD, IA, ZSL, IA e como sua utilização dos mesmos impacta o ramo da saúde, predição de dados e DM. A bibliografia deste trabalho apoia-se em trabalhos recentes e relevantes, compondo o estado da arte do tema explorado. Após este levantamento geral dos temas abordados, foram selecionados estudos de maior relevância como trabalhos correlatos.

Posteriormente, foram determinadas as bases de dados utilizadas para a implementação, seguido da limpeza e preparação dos dados. Então a escolha de quais técnicas e algoritmos foram aplicados a fim de predizer os dados e compará-los ao ZSL.

Por fim, realizados testes e adequações a fim de garantir a precisão dos resultados obtidos, foi observada a existência de pontos nos quais podem-se destacar benefícios da utilização do ZSL em contrapartida dos meios convencionais para classificação de dados que se baseiam em métricas selecionadas, principalmente acurácia da classificação e tempo de processamento obtidos.

1.4 Organização da Monografia

Este primeiro capítulo possui um teor introdutório sobre a realidade e benefícios de se usufruir de ferramentas como a IA no escopo da saúde, de modo a conter a motivação e designar um objetivo claro a ser alcançado com a realização deste trabalho. Os demais capítulos desta monográfica são compostos pelos seguintes conteúdos:

a) Capítulo 2 – Revisão Bibliográfica - São revisados conceitos do estado da arte de IA's na área da saúde, bem como conceitos aprofundados do ZSL e demais técnicas a serem utilizadas para a classificação dos dados. Também são apresentadas as definições das métricas pretendidas para avaliação posterior.
b) Capítulo 3- Metodologia e Desenvolvimento do Trabalho - É apresentada a descrição do processo seguido no desenvolvimento do trabalho.
c) Capítulo 4- Validação do trabalho realizado: apresentação dos resultados obtidos e análises realizadas.
d) Capítulo 5- Conclusão: apresentação da contribuição científica bem como as conclusões deste trabalho e sugestões de trabalhos futuros.

2 Fundamentação Teórica

2.1 Considerações Iniciais

Neste capítulo é apresentada a fundamentação teórica que serviu de alicerce para a concretização deste trabalho. Será seguida a seguinte apresentação dos conceitos: definição e conceituação de *Big Data*, KDD, *Data Mining*, Pré-processamento — processo que engloba a etapa de *Data Cleaning* —, IA, Predição de Dados — além da conceituação, a aplicação da mesma na área da saúde — e por fim os trabalhos correlatos encontrados na literatura científica.

2.2 Big Data

Presume-se que a palavra “dados” começou a ser usada por volta do século XVII, derivada do latim que significa “fatos dados ou concedidos”. O conceito de dados como algo dado foi criticado por muitos cientistas sociais, que afirmaram que o conceito simboliza, na verdade, “os fatos tomados ou observados da natureza pelo cientista” e deveria ser caracterizado como “capta”, que significa tomado e construído (ABADI, 2021; KITCHIN, 2014; DRUCKER, 2011). Ainda, de acordo com Abadi (2021), o século XX marca a evolução moderna do conceito de “dados” e “algoritmos”, no qual dados agora assumem uma forma digital e se referem a coleções de elementos binários transmissíveis e armazenados por cada operação computacional — isto é, a implementação dos algoritmos — que ocorre.

Nos últimos anos, pode-se observar uma demanda crescente por soluções que ofereçam ferramentas analíticas eficazes. Essa tendência é perceptível na análise de grandes volumes de

dados (BATKO; ŚLĘZAK, 2022). O conceito de *Big Data* evoluiu embora não haja na literatura uma definição exata para o mesmo. Entretanto, apesar das diversas definições, o *Big Data* pode ser tratado como uma grande quantidade de dados digitais, grandes conjuntos de dados, ferramenta, tecnologia ou fenômeno, seja cultural ou tecnológico (BATKO; ŚLĘZAK, 2022). Uma vez que o conceito alude predominantemente a um fenômeno que ocorre ao longo do tempo e não a uma tecnologia propriamente dita, diversos autores o descrevem segundo características que assimilam uma coleção de Vs relacionados à natureza do mesmo (BATKO; ŚLĘZAK, 2022; AGRAWAL; CLOUDHAY, 2019; AL MAYAHI, AL-BADI, TARHINI, 2018; FANG et al., 2015; GANDOMI; HAIDER, 2015; OLSZAK; MACH-KRÓI, 2018):

- Volume - refere-se à quantidade de dados, o maior desafio na análise *Big Data*,
- Velocidade - diz respeito à velocidade com a qual novos dados são gerados, pois o desafio é gerenciar os dados de forma efetiva em tempo real,
- Variedade - indica diferentes tipos de dados em um conjunto — principalmente na saúde —, cujo desafio é entregar perspectivas ao analisar todos os tipos de dados disponíveis de maneira integral,
- Variabilidade - trata-se da inconsistência encontrada nos dados, a qual exige correta interpretação de quais dados podem variar de forma significativa dado o contexto,
- Veracidade - refere-se à confiabilidade dos dados e sua qualidade,
- Visualização - relaciona-se à interpretação dos dados e perspectivas geradas a partir dos mesmos, dificultada pelas características supracitadas,
- Valor - o objetivo da análise do *Big Data* é descobrir conhecimento ocultado em grandes quantidades de informações.

O advento do *Big Data* e seu uso generalizado em registros eletrônicos de saúde para pacientes, permitiram a busca por soluções para questões de saúde de uma população, consideradas impossíveis anteriormente. Isto é possível, pois ao invés de extrapolar dados obtidos a partir de um pequeno número de amostras e fazer inferências, agora é possível utilizar-se de dados clínicos em nível populacional para obter uma visão mais precisa e real do mundo (NGIAM; HOR, 2019). Assim, a indústria da saúde produz diariamente um grande montante de dados, os quais compõem um importante domínio de aplicação dentro do *Big Data*. Para que seja possível fornecer aos pacientes os melhores serviços e atendimento, instituições médicas de vários países propuseram modelos de sistemas de informação médica (YANG et al., 2020; RISTEVSKI; CHEN, 2018). Portanto, com a evolução das informações médicas, numerosos dados digitais foram produzidos por meio de serviços médicos, cuidados e gestão da saúde que

formam assim o *Big Data* médico, o qual advém de várias fontes, como registros administrativos, registros clínicos, registros eletrônicos de saúde, relatórios de pacientes, biometria e mais. (YANG et al., 2020; LEE, YOON, 2017; OSSIO et all., 2017; RUMSFELD; JOYNT; MADDOX, 2016).

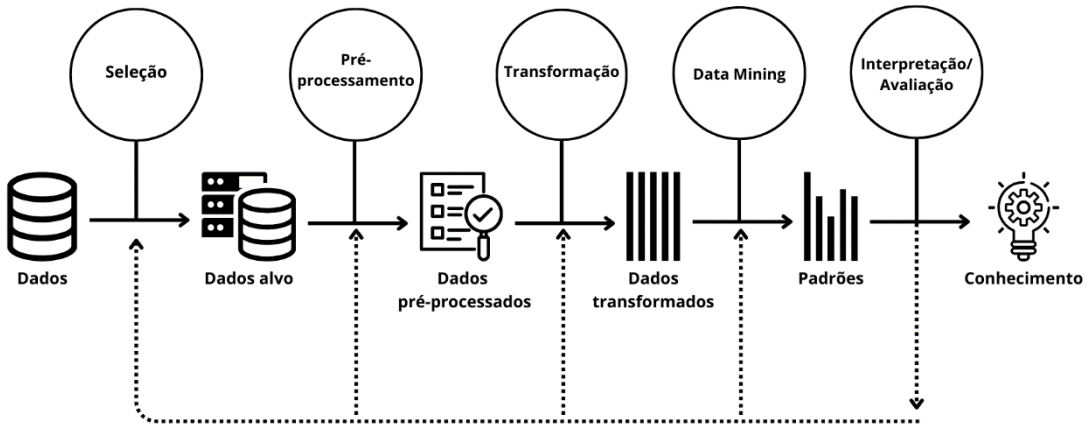
2.3 Extração de Conhecimento em Base de Dados

Há atualmente uma significativa quantidade de dados que são produzidos diariamente, porém é inviável ao ser humano analisar esses dados por si só. Claramente, uma pequena parcela de tais dados seria de fato visto por um olho humano, pois para serem definitivamente compreendidos teriam de ser analisados por computadores (FRAWLEY; PIATETSKY-SHAPIRO; MATHEUS, 1992). Assim, a comunidade científica da Computação tem a possibilidade de investir esforços no sentido de atender aos desafios científicos e práticos atrelados à extração de informação útil em meio a essa quantidade importante de dados. Chega-se assim ao termo Knowledge Discovery (Descoberta de Conhecimento) que se refere à extração não trivial de informação previamente desconhecida, implícita, e de possível utilidade de dados (FRAWLEY; PIATETSKY-SHAPIRO; MATHEUS, 1992).

O modelo de operação do KDD busca levar em consideração todas as etapas relevantes, desde acessar e selecionar os dados até à mineração dos dados propriamente dito. Data mining, é frequentemente empregado de maneira indispensável em conjunto com o KDD, porém é apenas uma das etapas do processo completo (KAMM; JAZDI; WEYRICH, 2021). Afirmar que o KDD é um processo sugere que o mesmo possua várias etapas, as quais envolvem a preparação dos dados, busca por padrões, avaliação do conhecimento e refinamento, repetidas até chegar-se a um resultado. Por não trivial deve-se entender que há uma busca ou inferência envolvida, ou seja, não é como um cálculo matemático direto no qual se tem conhecimento das variáveis predefinidas (KAMM; JAZDI; WEYRICH, 2021). O processo de KDD pode ser sumarizado nos seguintes passos: 1) primeiramente deve-se entender o domínio da aplicação e a prioridade da relevância do conhecimento, além de identificar um objetivo de ponto de vista do cliente; 2) criar um conjunto de dados alvo, por meio da seleção de um conjunto ou amostragem de conjuntos existentes, nos quais a descoberta será feita; 3) realização da limpeza dos dados ou *Data Cleaning* e pré-processamento — isto é, remoção de ruídos quando necessário, discernir estratégias para lidar com dados faltantes, entre outras ações; 4) reduzir a quantidade e escopo dos dados, seja a partir de métodos para a redução da dimensionalidade ou

transformação, o número efetivo de variáveis pode ser reduzido, ou por meio de representações invariantes para os dados podem ser encontrados; 5) encontrar métodos particulares de *Data Mining* que correspondam aos objetivos do processo de KDD; 6) realizar a análise exploratória seleção de modelo e hipótese, ou seja, escolher os algoritmos de *Data Mining* e selecionar os métodos que serão utilizados para a busca por padrões; 7) é o *Data Mining* em si, a busca por padrões de interesse; 8) interpretação dos padrões minerados que provavelmente retornará a alguma etapa intermediária entre 1 e 7 para reiteração; 9) Por fim, atuar no conhecimento descoberto, isto é, utilizar o conhecimento diretamente ao inclui-lo em um sistema para ação futura ou apenas ao documenta-lo e reporta-lo as partes interessadas (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996). Na figura 1 está ilustrada uma versão resumida desse processo.

Figura 1 - Processo de KDD



Fonte: Adaptado de (KAMM; JAZDI; WEYRICH, 2021)

2.4 Pré-processamento

Conforme mencionado na seção anterior, é muito provável que os dados em primeira instância estejam em um estado inoperável. Tais imperfeições nos dados podem variar de inconsistências e redundância que tornam o processo de data mining inviável um primeiro momento. Quanto maior for a quantidade de dados coletadas, maior a necessidade de

mecanismos sofisticados para analisá-los (GARCIA et al., 2016). O pré-processamento é capaz de adaptar os dados aos requisitos desejados por cada algoritmo de *data mining*, de modo a processar os dados inviáveis — que provavelmente seriam descartados — de uma forma diferente (GARCIA et al., 2016).

Contudo, apesar de ser uma ferramenta robusta que pode possibilitar que o usuário trate dados complexos que de outra forma seriam descartados e que é uma etapa quase inevitável atualmente, o pré-processamento pode provar-se custoso computacionalmente em termos de tempo de execução (GARCIA et al., 2016). O pré-processamento pode ser divido em etapas por si só, as quais são:

- Transformação - transformar os dados do jeito que serão utilizados pelos métodos de data mining;
- Integração - junção dos dados em um único conjunto de dados;
- Limpeza - tratamento de anomalias nos dados e normalização dos dados, padronização dos dados a fim de seguirem uma mesma regra.

Somente após uma etapa de pré-processamento bem-sucedida se pode aplicar qualquer algoritmo de data mining com uma fonte confiável de dados (GARCIA et al., 2016). Neste trabalho será dado um enfoque no passo de *Data Cleaning*, vista sua importância mediante o pré-processamento.

2.4.1 Limpeza de Dados

Conforme mencionado nas seções anteriores, o processo de *Data Cleaning* trata-se de um processo que se utiliza de estratégias para aumentar a qualidade dos dados a serem analisados. Tendo em vista a importância de cada dado para melhorar a tomada de decisão de quem os analisa — inclusive para IA’s —, a qualidade dos dados se mantém um tópico de extrema relevância, uma vez que dados “sujos” podem levar a decisões incorretas e análises não confiáveis (CHU et al., 2016). Alguns exemplos de ruídos nos dados seriam dados faltantes, erros de digitação, formatos misturados, duplicatas de um mesmo registro e violações de regras de negócio. Ainda, pode-se dividir as técnicas de *Data Cleaning* em duas classes, as quantitativas: aplicadas para a detecção de outliers — valores significativamente discrepantes dos demais do conjunto de dados —, por exemplo um salário com três desvios padrões do salário médio é um erro a ser corrigido; as qualitativas que utilizam restrições, regras e padrões, por exemplo não é permitida a existência de dois funcionários que ocupem a mesma posição na

hierarquia de uma empresa, mas possuam salários diferentes por não residirem no mesmo local (CHU et al., 2016).

O entendimento do *Data Cleaning* poderia ser aprimorado ao ser dividido nas seguintes etapas (RAHM; DO, 2000):

- Análise dos dados: para conhecer quais tipos de anomalias estão presentes e precisam ser removidas.
- Definição do fluxo de trabalho e regras de mapeamento: a depender do número de fontes de dados a serem integradas, pode ocorrer uma "heterogeneidade de sujeiras", ou seja, a mistura de dados com diferentes tipos de ruído.
- Verificação: a correção e eficácia da transformação devem ser testadas e avaliadas e, se necessário, adequar as definições de transformação.
- Transformação: execução da fase de transformação.
- Backflow (retorno ou retrocesso) dos dados limpos: após uma fonte de dados ser limpa, tais dados devem também ser substituídos na fonte original a fim de melhorar aplicações legado e evitar refazer a limpeza dos dados em futuras extrações.

2.5 Mineração de Dados

O conceito de *Data* Mining, em português, Mineração de Dados, tem sua origem fortemente ligada a campos como a estatística, IA, ML e bancos de dados. Em suma, o *Data Mining* trata-se de uma etapa do processo de KDD, que consiste em aplicar algoritmos de análise e descoberta de dados com o intuito de identificar padrões (ou modelos) (MIKUT; REISCHL, 2011). Como o processo de *Data Mining* é a etapa que irá levar a descoberta de conhecimento dentro do KDD, por vezes pode ser referenciado como sinônimo da descoberta de conhecimento — apesar de ser uma etapa e não o processo por completo —. Dessa forma, nos últimos 10 a 15 anos, *Data mining* tem se tornado uma tecnologia por si só, é bem estabelecida na área de inteligência dos negócios e continua a demonstrar uma importância cada vez maior e constante nos setores de tecnologia e ciências da vida (MIKUT; REISCHL, 2011).

As técnicas do *Data Mining*, estão classificadas em: classificação, regressão e agrupamento (ou clustering). A utilização destas técnicas de mineração de dados depende das características dos dados e das percepções específicas procuradas (CHEN et al., 2023; JOTHI; HUSAIN, 2016; SZEGEDY et al., 2013).

2.6 Inteligência Artificial

A criação de máquinas pela humanidade sempre almejou em facilitar uma tarefa, um exemplo disto é a execução de diversos trabalhos intensos pela tecnologia atual (JIANG et al., 2022). Entretando, em muitas ocasiões, em virtude da demanda por maior produtividade ou até mesmo pela simples curiosidade da descoberta, tenta-se cada vez mais replicar o comportamento da inteligência humana e seu processo de aprendizado em máquinas, o que originou a motivação para a IA (JIANG et al., 2022).

Existem múltiplas definições para IA, como por exemplo a do teste de Turing, no qual IA é definida como máquinas com a habilidade de se comunicar com humanos sem revelar sua identidade como não humanos, de forma que o julgamento — ser ou não humano — é binário (JIANG et al., 2022). Porém, independentemente das diversas definições na literatura, atualmente o conceito de IA é amplamente aceito como as teorias de pesquisa, métodos, tecnologias e aplicações para simular, estender e expandir a inteligência humana (JIANG et al., 2022). Dessa forma, assim como o vapor impulsionou a era industrial, a IA impulsiona a era da tecnologia na contemporaneidade e além dela (JIANG et al., 2022).

A ascensão da IA se deve a alguns fatores históricos, primeiramente, o sucesso do ML no final dos anos 1980 tem fomentado teorias como árvores de decisão, máquinas de veto de suporte, Adaboost — algoritmo de ML — e florestas aleatórias (JIANG et al., 2022). Após, o o acesso a grandes quantidades de informações com o *Big Data* — citado nas seções anteriores — trouxe uma chance ímpar no treinamento de modelos eficientes para previsões e classificações que incentivam e corroboram para que a comunidade científica construa bancos de dados excepcionais (JIANG et al., 2022). Além disso, o desenvolvimento de hardwares mais capazes proporcionou que máquinas baseadas em IA lidem com tarefas mais complexas (JIANG et al., 2022).

No geral, IA’s operam de três maneiras principais, com aprendizado supervisionado, não supervisionado e aprendizado por reforço (JIANG et al., 2022). No primeiro caso, ao invés de especificar as entradas e os algoritmos definidos por humanos para se chegar nos resultados, a alternativa é selecionar as entradas e seus resultados esperados para que a IA (máquina) encontre automaticamente padrões que geram assim um aprendizado (JIANG et al., 2022; SZEGEDY et al., 2013; ZHANG et al., 2021). No segundo caso, como os rótulos de treinamento não estão disponíveis, abordagens baseadas ema agrupamentos — denominados clusters — podem dividir os dados em grupos sem especificar o mapeamento exato para o espaço de características. Por fim, o aprendizado por reforço usa a ideia da IA aprender

continuamente, assim ações e se ajustam as regras de atuação ajustadas com base em avaliações de feedback ou recompensas. (JIANG et al., 2022; MATIGNON; LAURENT; LE FORT-PIAT, 2007).

2.7 Predição de Dados

O campo do ML é apenas um subconjunto do que forma a IA, subconjunto este que objetiva desenvolver algoritmos preditivos embasados na ideia de que as máquinas devem ser capazes de acessar dados e aprender por conta própria (BADAWY; RAMADAN; HEFNY, 2023). A predição de dados é o processo de usar métodos de ML para fazer predições ou previsões sobre eventos futuros, útil em aplicações como medicina (ATHEY, 2017). Assim como Krishnamoorthi et al. (2022) demonstraram ao utilizar dados sobre diabetes para predição de predisposição a doença com uso de um aprendizado supervisionado, primeiro ocorre a separação dos dados em um conjunto de dados para o treinamento e o conjunto alvo. Assim, algoritmos de ML são treinados e depois expostos aos dados do conjunto alvo para classificação.

2.7.1 Predição de Dados na Saúde

A utilização de IA's na área da saúde possui grande potencial como já apresentado em seções anteriores deste trabalho. Os tipos de IA podem variar de acordo com a necessidade, mas em sua maioria usam ML ou processamento de sinais combinados com paradigmas de aprendizado que geram para a saúde: diagnósticos, avaliações de risco de morbidade ou mortalidade de pacientes, previsão e vigilância de surtos de doenças e políticas e planejamento de saúde (SCHWALBE; WAHL, 2020). Assim como Boudreaux et al. (2021) apresentaram, a predição de dados com a utilização de IA em meio a campos da saúde pode não só prever doenças físicas, mas também quadros clínicos de distúrbios mentais que ocasionariam na morte do paciente.

2.7.2 Métodos Convencionais de Predição

Esta sessão é reservada para a apresentação de alguns algoritmos e métodos mais comumente utilizados para a predizer dados. Posteriormente, os mesmos serão implementados e comparados com o a técnica alvo deste trabalho, o ZSL.

2.7.2.1 K-ésimo Vizinho Mais Próximo

O K-nearest neighbor (KNN) é um algoritmo voltado para problemas de classificação onde para antecipar o rótulo alvo do conjunto de dados teste, o KNN determina a distância dos rótulos de classe dos dados de treinamento mais próximos com um novo ponto de dados de teste, na presença de um valor K, em seguida, calcula o número de pontos de dados mais próximos através do valor K e determina o rótulo da nova classe de dados de teste (BADAWY,; RAMADAN; HEFNY, 2023).

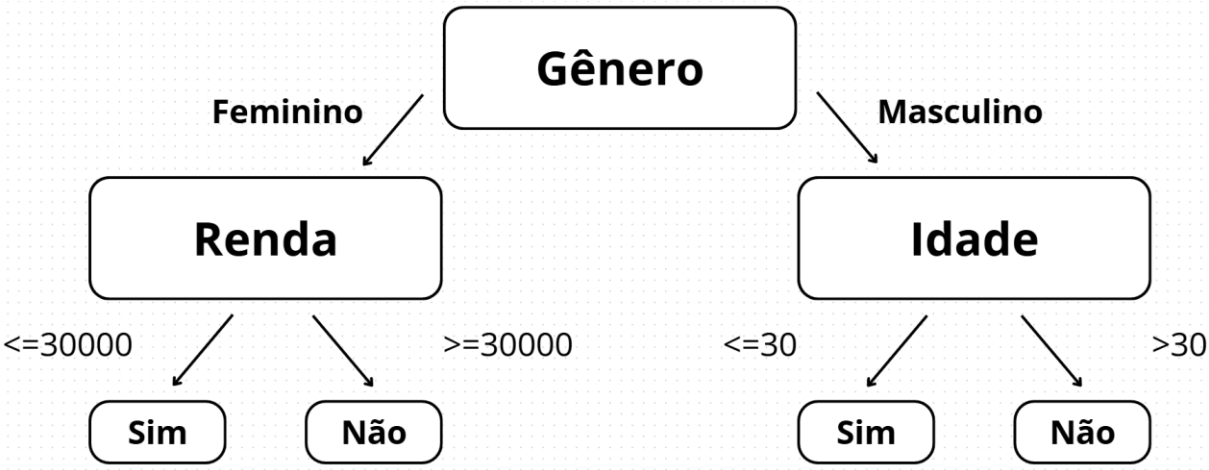
O KNN apresenta muitas vantagens como: é suficientemente poderoso se o tamanho da base de dados de treinamento for grande, é simples e flexível com atributos e funções de distância, pode lidar com conjuntos de dados multiclasse (BADAWY; RAMADAN; HEFNY, 2023). Da mesma forma, apresenta desvantagens também como: dificuldade em escolher um valor adequado de K, a dificuldade na escolha do tipo de função de distância para um conjunto de dados específico e o custo computacional um pouco elevado devido à distância entre todos os pontos de dados de treinamento (BADAWY; RAMADAN; HEFNY, 2023).

2.7.2.2 Árvore de Decisão

A árvore de decisão é um método de aprendizado supervisionado destinado a classificações, onde combina os valores de atributos baseado em sua ordem, seja ascendente ou descendente (BADAWY; RAMADAN; HEFNY, 2023; DHALL; KAUR; JUNEJA, 2020). A árvore de decisão determina cada caminho a partir da raiz ao usar uma sequência de separação de dados até atingir uma conclusão booleana no nó folha (BADAWY; RAMADAN; HEFNY, 2023; YANG, 2019).

Árvores de decisão podem apresentar desvantagens como aumento da complexidade com a nomenclatura crescente, pequenas modificações podem levar a uma arquitetura diferente, além de levar mais tempo no processamento dos dados no treinamento (BADAWY; RAMADAN; HEFNY, 2023; GUPTA; PANDYA; 2022). Na figura 2 está ilustrado um exemplo da árvore de decisão.

Figura 2 : Exemplo de Árvore de Decisão



Fonte: Adaptado de (BADAWY; RAMADAN; HEFNY, 2023)

2.7.2.3 Floresta Aleatória

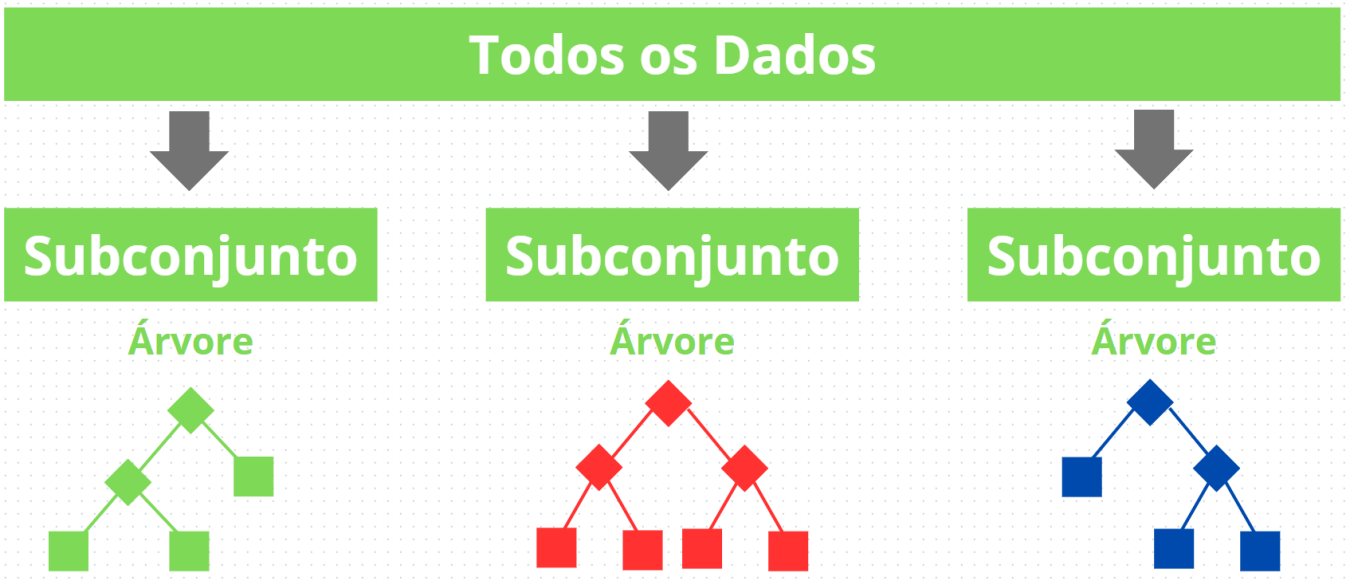
Floresta aleatória ou Random Forest é uma técnica básica que produz resultados corretos na maioria das vezes, tanto para regressão quanto para classificação. A ideia é produzir um conjunto de árvores de decisão e combiná-las ((BADAWY; RAMADAN; HEFNY, 2023; BAKYARANI et al., 2019).

Dessa forma, quanto maior o número de árvores de decisão inseridas na floresta, maior a acurácia dos resultados, pois cada árvore de decisão é criada apenas com parte do conjunto de dados original e treinada com aproximações (BADAWY; RAMADAN; HEFNY, 2023). Assim, o método seleciona um subconjunto de características dos dados e para cada subconjunto gera n árvores aleatórias, e posteriormente combina os resultados das árvores de decisão e fornece uma saída final (BADAWY; RAMADAN; HEFNY, 2023; SRIVASTAVA; SAINI; GUPTA, 2019). A floresta aleatória trabalha com dados rotulados para realizar predições e construir um modelo e, ao final do processo, o modelo é utilizado para classificar dados que ainda não possuem rótulos atribuídos (BADAWY; RAMADAN; HEFNY, 2023).

Por fim, a técnica apresenta vantagens, como poder ser utilizada para determinar a relevância das variáveis em uma tarefa de regressão ou classificação (BADAWY; RAMADAN; HEFNY, 2023; HO, 1998; HOFMANN; KLINKENBERG, 2016). Tal relevância é definida por uma escala, baseada na impureza atribuída a cada nó usado para a segmentação dos dados (BADAWY; RAMADAN; HEFNY, 2023; CHOW; LIU, 1968). Também, automatiza os

valores ausente contidos nos dados e resolve o problema de overfitting das árvores de decisão, além de poder lidar eficientemente com grandes conjuntos de dados. Entretanto, sofre de desvantagens como a necessidade de mais poder computacional e recursos para gerar os resultados e mais esforço de treinamento devido as múltiplas árvores de decisão (BADAWY; RAMADAN; HEFNY, 2023). Na figura 3 está ilustrada um exemplo da árvore aleatória.

Figura 3: Exemplo de Floresta Aletória



Fonte: Adaptado de (BADAWY; RAMADAN; HEFNY, 2023)

2.7.3 Aprendizado Zero-Shot (zero amostras)

O ZSL é um assunto emergente nos últimos anos. O método visa distinguir imagens de classes não conhecidas utilizando imagens de classes vistas para treinar o classificador, no qual a informação de discriminação geralmente está contida nas imagens como por exemplo as listras pretas e brancas de uma zebra são a principal diferença em relação a um cavalo. (XIE et al., 2022).

Dessa forma, no ZSL os dados de treinamento para classes de teste não vistas não estão disponíveis e o conjunto de rótulos para os dados de treinamento e teste são disjuntos (XIE et al., 2022). Por conseguinte, o ZSL demonstra capacidade inerentes de lidar com questões de rotulagem de dados e reconhecimento de imagens não vistas (XIE et al., 2022). O que torna o ZSL viável é a garantia de uma transferência de conhecimento eficaz através das classes

conhecidas para as desconhecidas, a descrição textual de cada classe é essencial para garantir as categorias vistas e as não vistas (XIE et al., 2022). Um pipeline típico do processo de ZSL é o modelo de incorporação, o qual aprende a discernir pesos — por meio de conexão de características espaciais e semânticas — baseado nas classes e atributos correspondentes (XIE et al., 2022).

Portanto, é um paradigma de aprendizado que pode ser comparado ao comportamento de aprendizado humano, pois ao observar algo que não se conhece, procura-se uma aproximação baseado em conhecimentos prévios adquiridos (POURPANAH et al., 2023; ROMERA-PAREDES; TORR, 2015).

2.8 Trabalhos Correlatos

Os trabalhos que visam a utilizam a IA na esfera da saúde são inúmeros, porém como o estudo e aperfeiçoamento da IA é constante, há o constante surgimento de uma profusão de estudos. Para exibir uma pequena amostra do potencial uso de IA juntamente com o ZSL na saúde serão examinados os trabalhos de (BHATE et al., 2023), (WORKMAN et al., 2023), (FARHAD et al., 2023) e (KAUFMANN et al., 2023).

2.9 Aprendizado Zero-shot com Instrução Mínima para Extração de Determinantes Sociais e Histórico Familiar de Notas Clínicas usando Modelo GPT

O trabalho proposto por Bhate et al. (2023) demonstra a aplicação da abordagem ZSL juntamente com o uso da tecnologia do modelo GPT. No ramo da saúde, informações não estruturadas acabam por ser comuns, tais como determinantes sociais e histórico familiar e há um esforço de pesquisa que objetiva deixar tais exemplos estruturados visto que podem fornecer melhores resultados. No ZSL, informação semântica é um recurso precioso uma vez que é baseada nela que o modelo fará a classificação das classes não vistas. Dessa forma, o trabalho propõe uma investigação ao utilizar o ZSL para a extração de informações fornecendo informações mínimas ao modelo GPT.

Assim, os dados para a realização deste trabalho foram extraídos de 1.000 notas clínicas narrativas de mais de 150 pacientes, todas anomizadas, ou seja, toda e qualquer informação pessoal presente nas notas foram retiradas. Então, a rotulagem social e de histórico familiar foi

feita por dois rotuladores (ou anotadores) independentes. Um terceiro ficou encarregado de resolver casos conflitantes entre os dois principais.

O uso da API OpenAI GPT-3.5 ocorreu com dois papéis distintos, o de sistema e o de usuário. O sistema foi utilizado para fornecer as instruções da tarefa para que o modelo atuasse como um anotador, enquanto o usuário forneceu um texto de entrada para o ZSL. No final, um prompt simples é montado, com o sistema que informou ao modelo como se comportar com informações mínimas de forma a processar a entrada e fornecer os resultados em um formato dado, o qual inclui os tipos a serem extraídos.

Um pré-processamento foi aplicado com o intuito de denotar os subtipos do histórico social e familiar. Assim, se uma substring tiver sido extraída como história social, ela seria subdividida cm base em pontuações como vírgulas e demais pontuações e cada segmento, palavras-chave específicas de subtipos categorizaram em subtipos. O exemplo dado no trabalho é a string “desempregado, sem histórico de tabagismo, algum ensino superior” seria subdividida em “desempregado”, “sem histórico de tabagismo” e “algum ensino superior”. Palavras como “escola”, “faculdade”, “universidade” e outras foram usadas para determinar se um segmento é relevante para status educacional.

Com isso, as métricas de avaliação utilizadas foram as métricas de Reconhecimento de Entidades Nomeadas (NER) juntamente com avaliações semânticas. Para o NER, foi utilizada correspondência relaxada, ou seja, todos os conceitos extraídos foram convertidos em palavras. Então, palavras que estão na rotulagem e na saída do GPT são consideradas Verdadeiros Positivos, as que estão na saída do GPT e não na rotulagem são Falsos Positivos e as que estão na rotulagem e não na saída do GPT são Falsos Negativos. Já para a avaliação semântica, a rotulagem e o conceito extraído foram convertidos em representações vetoriais e sua similaridade foi medida com cosseno, se o limiar θ estivesse acima de 0,8, há correspondência e acima de 0,9 há uma correspondência exata.

Por fim, foram alcançados resultados satisfatórios onde foi mostrado que com instrução mínima, o modelo GPT atingiu grande valor de métrica F1 em extração de informação demográfica com 0.975, moderada em extração de determinante social (0.615) e de histórico familiar (0.722).

2.10 Identificando Documentação de Suicídio em Notas Clínicas com Aprendizado Zero-Shot

Segundo o trabalho de Workman et al. (2023), a taxa de suicídio de veteranos de guerra nos Estados Unidos da América (EUA) tem se tornado um problema que exige atenção dado o seu constante crescimento. A utilização de novas abordagens como o ZSL é justificada, pois em estratégias como *Deep Learning*, existem poucas amostras de treinamento dado a pequena parcela de verdadeiros positivos em meio a um número de pacientes. Dessa forma, a utilização do ZSL a um contexto de classes não vistas baseado em classes vistas é bem-vindo.

O objetivo em si era uma classificação binária de possibilidade de suicídio, ou seja, suicida e não suicida. No auxílio do ZSL o conjunto de dados para treino foi criado a partir de códigos de diagnóstico. Uma string que representava o possível suicídio foi selecionada, em seguida foi construído o espaço semântico com características-chave associadas ao suicídio para o treinamento.

O conjunto de dados de treinamento foi construído a partir de dois conjuntos, onde o primeiro consistia em 50.000 notas clínicas aleatoriamente selecionadas do departamento de VA (Veteran Affairs, ou Assuntos dos Veteranos) de 2016 a 2019 que continham a string base “suicid” (o que engloba “suicide” e “suicidal”) e estavam associadas a ao menos um código de diagnóstico do Relatório Nacional de Estatísticas de Saúde dos Centros de Controle e Prevenção (CDC). O segundo conjunto consistia em 50.000 notas clínicas aleatoriamente selecionadas do VA de encontros laboratoriais do mesmo período que estavam relacionados a outros códigos de diagnóstico que não indicavam suicídio ou autolesão. Cada um dos dois conjuntos foi pré-processado da seguinte forma: todas as letras convertidas em letras minúsculas, remoção de marcações básicas de formatação e pontuação, separação de cadeias de caracteres em token (palavras), na qual *tokens* relevantes foram concatenados e houve a remoção de tokens que não fossem compostos inteiramente de letras.

A construção do espaço semântico foi feita em quatro etapas onde, respectivamente, foi identificada a lista de características que podiam ser relevantes para o treinamento de uma rotulagem positiva, foram criados palavras de incorporação (representações numéricas de objetos do mundo real) usando uma arquitetura skip-gram (modelo que dada uma palavras, tenta prever quais são palavras que aparecem no mesmo contexto), a partir das palavras de incorporação foram identificadas palavras de contexto e por último um peso contextual foi atribuído a cada característica para cada documento no mapeamento semântico do espaço para o espaço de características.

Então, os resultados apontaram que o ZSL identificou de forma efetiva a chance de suicídio em todos os tipos de notas clínicas e documentações de risco de suicídio de notas

clínicas não associadas a suicídio com o uso de códigos de diagnósticos ICD-10-CM (International Classification Disease, décima revisão, Modificação Clínica) com alta precisão.

2.11 Uma Abordagem Siamesa de Zero-shot e Few-shot Eficiente em Dados para Diagnóstico Automatizado de Hipertrofia Ventricular Esquerda

O trabalho de Farhad et al. (2023) trata da utilização de abordagens conjuntas de ZSL e FSL (Few-Show Learning) para o diagnóstico de Hipertrofia Ventricular Esquerda (HVE) que é um grande indicador de outras condições relacionadas ao funcionamento do coração. A motivação do trabalho ocorreu devido a disponibilidade de dados médicos devido a regras de privacidade de dados e leis. Assim, com seu próprio conjunto de dados, utilizaram as abordagens ZSL e FSL baseados em uma rede siamesa para classificar o HVE em imagens.

A criação do conjunto de dados para a classificação incluiu ecocardiogramas, contendo imagens de HVE de 2CH (nomenclatura para chambers, ou seja, 2 ventrículos) e 4CH (mesma nomenclatura para 4 ventrículos). A composição se deu por 18 imagens normais e 20 com HVE de 2CH e 18 normais e 14 com HVE de 4CH. Além disso, dois especialistas ecocardiógrafos conduziram as análises das imagens com HVE e as normais. Apenas as imagens consideradas pelos dois especialistas foram levadas para o treinamento do modelo.

Dessa forma, foi desenvolvido o CardiacSiamese-CAMZ (Cardiac Apical Multiview Zero-Shot), uma nova abordagem de ZSL específica para imagens HVE e normais da ecocardiografia, uma vez que as imagens médicas não possuem as mesmas cores e qualidades de imagens normais. Dado que há apenas uma classe para treinamento, cria-se um problema pois não haverá classe não similar para o treinamento do modelo e não há vetores de texto para avaliar as imagens médicas, pois são as mesmas em termos de forma e cor. Assim, par ao CardiacSiamese-CAMZ o treinamento consistiu apenas de imagens normais com dois pontos de vista (2CH e 4CH). O componente de extração de características converte uma imagem de entrada em u vetor de características unidimensional. Assim, após a extração de características foi calculada a distância euclidiana, assim, o processo de treinamento atualiza os pesos de modo que os pares de mesma classe fiquem mais próximos e vice-e-versa. Com o treinamento das imagens normais efetuado, o modelo calcula uma distância de corte que será o valor divisor entre as classes “Normal” e “HVE” durante o teste. Por fim, para o teste, uma imagem é selecionada do conjunto de dados e sua distância é calculada com três imagens aleatórias dos

dados de treinamento. A classe da imagem é atribuída após a comparação das distâncias entre as imagens de treino, de teste e a distância de corte.

Além disso, foi feito uma abordagem que foi denominada de CardiacSiamese Few-Shot learning onde o modelo foi treinado usando pares de imagens normais e HVE. Nesse processo, se as imagens são da mesma classe, são rotuladas par positivo e caso as duas sejam de classes diferentes, pares negativos. A extração de características de um par, são convertidas em vetores de características e a distância euclidiana entre elas é chamada de função de perda. Dessa forma, o treinamento atualiza os pesos para que a distância entre as mesmas classes diminua e a de classes diferentes aumente.

Em suma, os resultados alcançados demonstram que as classificações manuais dos especialistas — os quais foram escolhidos com expertise e mais de 10 anos de atuação na área — foi de alta consistência. Ainda, o modelo CardiacSiamese atingiu uma pontuação intra-observador mais elevada o que demonstrou que a expertise dos especialistas pode ser reproduzida pelo modelo.

2.12 Validação de Ferramenta de Processamento de Linguagem Natural com Zero-Shot-Learning para Facilitar a Abstração de Dados em Pesquisas Urológicas

O trabalho proposto por Kaufmann et al. (2023) trata de uma ferramenta de processamento de linguagem natural que utiliza ZSL, o qual alcança uma acurácia comparável à dos revisores humanos para tarefas de abstração de dados, oferecendo benefícios significativos em termos de economia de tempo. Contudo, a ferramenta requer refinamento e integração aos sistemas de prontuários eletrônicos dos hospitais para uso prático. O objetivo do estudo foi desenvolver e validar tal ferramenta de abstração de dados não estruturados em pesquisas urológicas e avaliar seu desempenho em relação a revisores humanos.

O estudo usou uma ferramenta de NPL com ZSL baseada no modelo GPT-3.5 da OpenAI, desenvolvida usando linguagem de programação python e diversas bibliotecas de terceiros. A avaliação foi feita com uso de 199 laudos patológicos de prostatectomia radical desidentificados, processados tanto em formatos vetorizados quanto digitalizados. Três revisores humanos foram utilizados para a comparação de desempenho com a ferramenta. Os resultados demonstraram que a ferramenta foi significativamente mais rápida, com um tempo médio de 12 segundos por relatório para relatórios vetorizados e 15 segundos para relatórios

digitalizados, em comparação com um tempo médio de 93 segundos por relatório dos humanos. Além disso, a acurácia da ferramenta não foi inferior, com uma acurácia de 94,2% para relatórios vetorizados e 88,7% para digitalizados.

As limitações do estudo incluem a possibilidade de erros introduzidos pelo reconhecimento óptico de caracteres, o que pode afetar o desempenho da ferramenta, além da necessidade de uma avaliação mais aprofundada do desempenho da ferramenta em diferentes ambientes clínicos. Ainda, apresenta dificuldades com: certas variáveis, as quais alcançaram acurácia na faixa de 83%, preocupações de privacidade, questões práticas para a obtenção e uso de uma chave de API da OpenAI, períodos de inatividade do servidor e desafios na formulação de consultas para obter os resultados desejados.

Dessa forma, o estudo concluiu que a ferramenta NPL com uso do modelo ZSL é um método altamente generalizável e preciso para abstração de dados de textos não estruturados em pesquisas urológicas, oferecendo uma solução promissora para os desafios de extração de dados de prontuários eletrônicos.

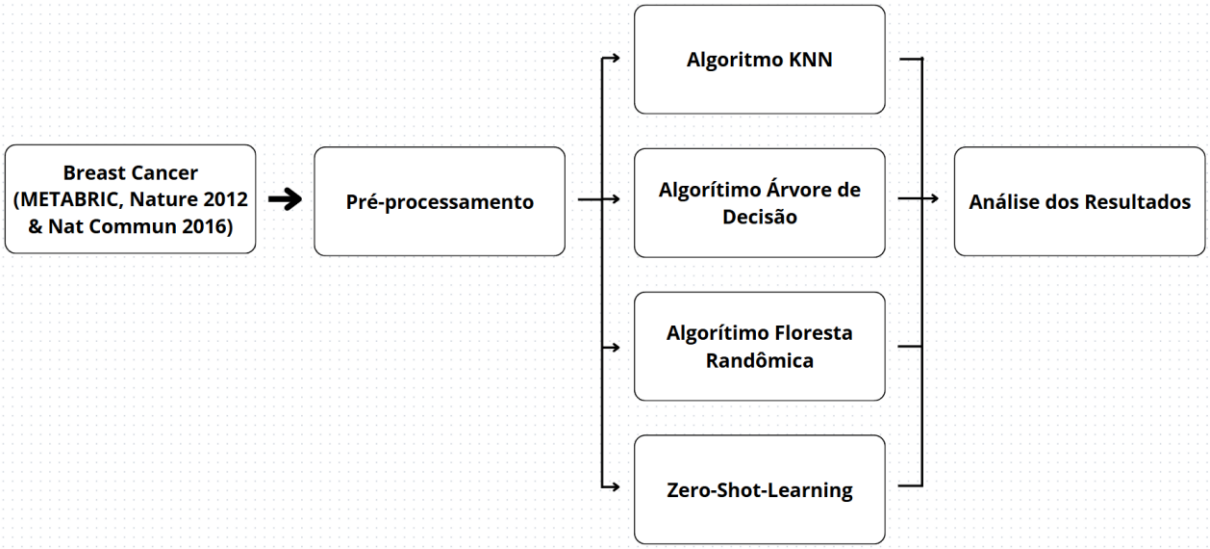
2.13 Considerações Finais

Com a descrição dos principais trabalhos da literatura, pode-se notar que o ZSL é uma abordagem recente e que exige aperfeiçoamento, porém com indicações de oferecer resultados mais interessantes do que as abordagens convencionais. Apesar de ter sido idealizado para dados relacionados a imagens, o ZSL pode ser adaptado para dados tabulares se aplicada as devidas correções como Workman et al. (2023) evidenciou.

3 Desenvolvimento

Neste capítulo são detalhados os processos utilizados para a avaliação dos métodos de predição de dados bem como a obtenção dos dados, pré-processamento e estratégias adotadas para originar os resultados alcançados. Para tal, são expostos fluxogramas e pseudocódigos a fim de facilitar a compreensão do processo. No fluxograma da figura 4 estão ilustradas as etapas, cujas descrições são apresentadas nas seções seguintes.

Figura 4 - Processos de Realização do Presente Trabalho



Fonte: Elaborado pelo autor

3.1 Obtenção dos Dados Analisados

Os dados utilizados neste trabalho foram originados do cBioPortal, um site público que foi gerado e desenvolvido no Memorial Sloan Kettering Cancer Center (MSK), o qual é hospedado pelo Centro de Oncologia Molecular do MSK.

A base escolhida foi Breast Cancer (METABRIC, Nature 2012 & Nat Commun 2016), mantida pelo cBioPortal, a qual reúne dados de câncer de mama. O conjunto contém mais de

2500 amostras de pacientes com câncer de mama, das quais têm-se informações como mutação somática, expressão de RNA, dentre outros dados clínicos detalhados. Algumas das informações são os atributos: ID do estudo, ID do paciente, ID da amostra, Idade do paciente quando diagnosticado, Tipo de cirurgia de mama, Tipo de câncer, Tipo de câncer detalhado, Celularidade, se o paciente fez quimioterapia, se o paciente fez hormonoterapia, dentre outras informações médicas mais precisas.

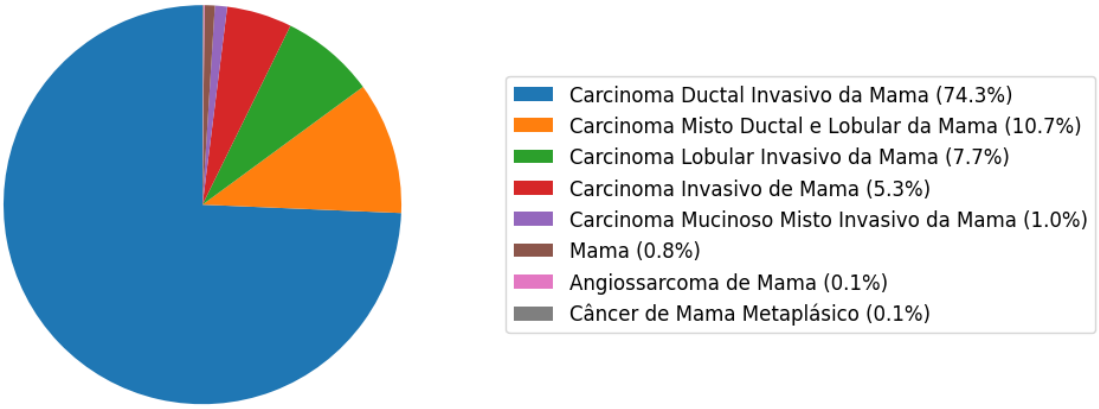
Além disso, este conjunto de dados já possui uso na bibliografia como no trabalho proposto por Dong, Pengzhi, et al. (2018), o qual teve finalidade identificar as principais vias e genes envolvidos do câncer de mama triplo-negativo e localizar potenciais mecanismos de início e progressão dele. Outro exemplo é o trabalho proposto por Blockhuys, Stéphanie, Donita C. Brady, e Pernilla Wittung-Stafshede (2020), o qual estudou a relação do cobre com o câncer de mama, dado o fato de que o mesmo causa migração de células de câncer de mama ocasionando a metástase, que é uma das principais causas de morte relacionadas ao câncer de mama.

3.2 Pré-processamento

A etapa de pré-processamento como citado no capítulo 2, é fundamental quando se analisa dados. Os dados foram obtidos diretamente no formato adequado para a importação dentro da aplicação do Google Colab, plataforma a qual será utilizada para a realização deste trabalho. A partir disso, os nomes das colunas foram formatados para um formato sem espaços, substituindo-os por “_”.

A predição será feita em cima da coluna que descreve o tipo detalhado de câncer de mama do paciente, sendo eles: câncer de mama, angiossarcoma de mama, carcinoma ductal invasivo de mama, carcinoma lobular invasivo de mama, carcinoma mucinoso invasivo misto de mama, carcinoma misto ductal e lobular de mama, carcinoma invasivo de mama e câncer de mama metaplásico. Dessa forma, para melhor visualização, na figura 5 é apresentado o gráfico gerado para exibir a distribuição de cada classe para os valores atribuídos em toda base de dados ainda não tratada.

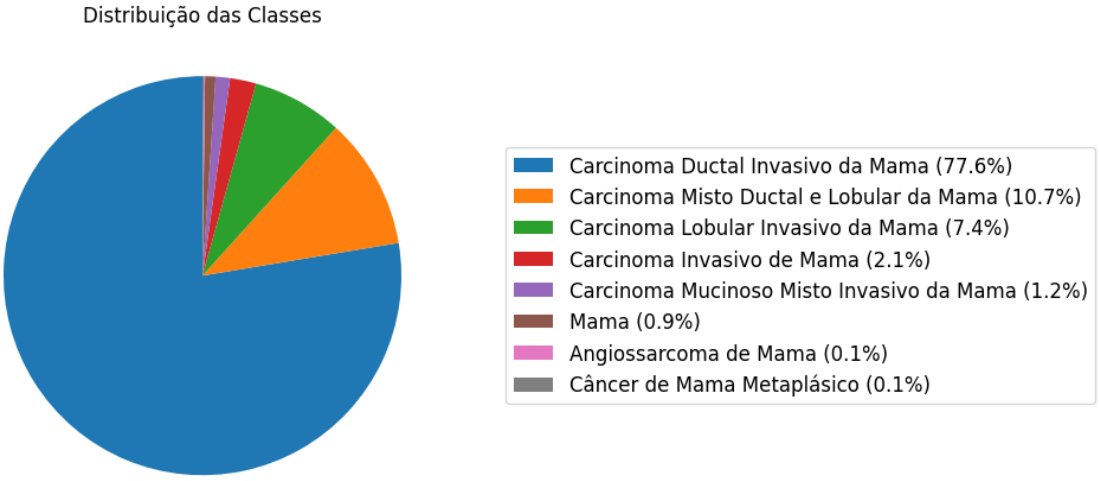
Figura 5 – Distribuição de Classes Antes do Pré-processamento



Fonte: Elaborado pelo autor

Após a avaliação da distribuição, registros da base com muitos valores nulos ou faltantes foram eliminados. Atributos identificadores como identificação do estudo, identificação do paciente e identificação da amostra também foram removidas tendo em vista que não agregariam a classificação. Também, registros que possuíssem valores nulos ou faltantes no atributo de tipo detalhado de câncer, o atributo alvo da classificação, foram removidos. Por fim, uma função foi definida para o preenchimento de valores faltantes ou nulos de cada coluna de acordo com o tipo de valor, cujo preenchimento de valor faltante foi efetuado pela moda em colunas que possuam valores categóricos e a média para colunas com valores numéricos, e que resultou num total de 1980 registros dos 2509 registros originais. Na figura 6 é apresentada a distribuição das classes após a base ser limpa de dados nulos.

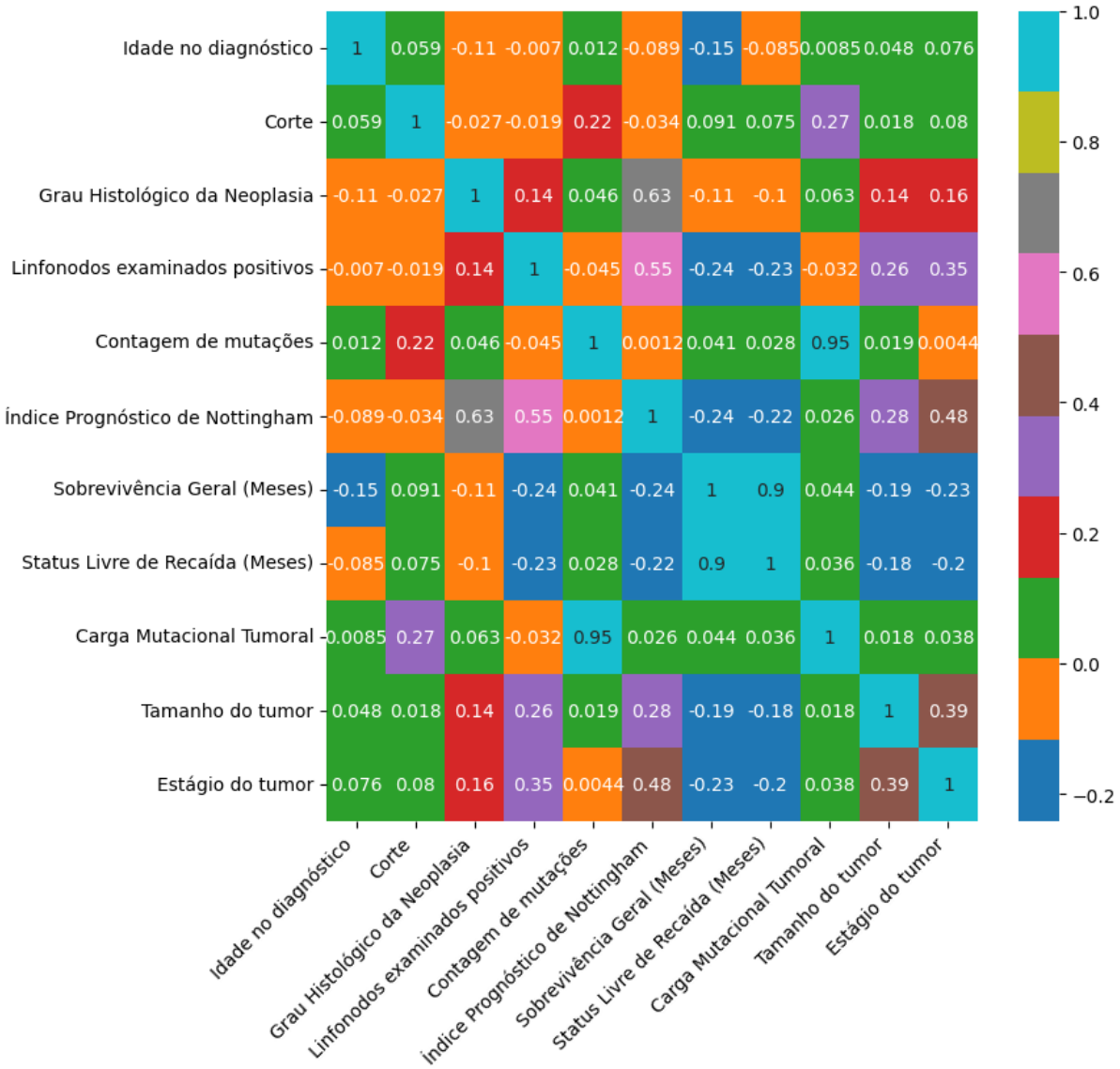
Figura 6 – Distribuição das Classes Após Remoção de Valores Nulos e Preenchimento de Valores Faltantes



Fonte: Elaborado pelo autor

Seguindo, foram identificados os atributos redundantes, isto é, atributos com alto grau de similaridade de importância entre si e podem ser removidos para diminuição da complexidade e ajuste excessivo para tal, a matriz de correlação da figura 7 foi feita, utilizou-se um limiar de correlação de 0.7 que resultou na remoção de dois atributos: carga mutacional do tumor (não sinônima) e estado livre de recaída (meses).

Figura 7 – Matriz de Correlação



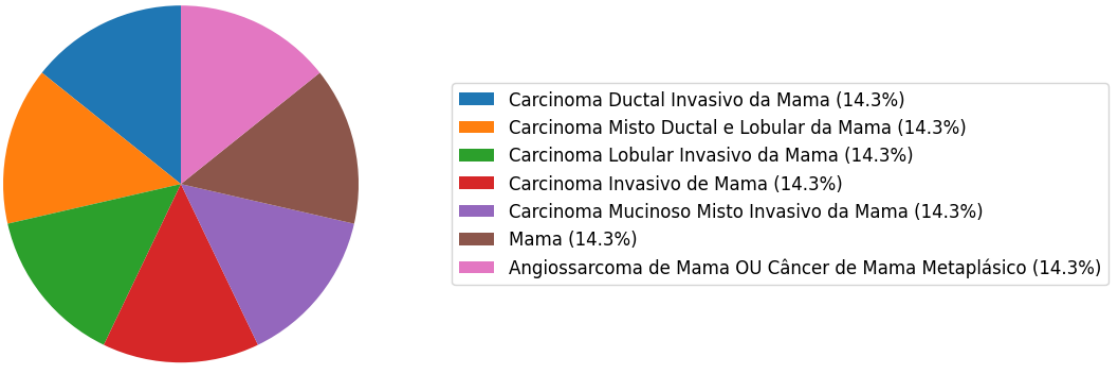
Fonte: Elaborado pelo autor

Além dessa seleção, houve uma segunda seleção de características baseada na importância de cada uma a partir da ferramenta de floresta randômica da biblioteca sklearn, que eliminou mais features desnecessárias ao modelo. No fim, as seguintes características foram selecionadas: Idade no diagnóstico, Grau Histológico da Neoplasia, Outro Subtipo Histológico do Tumor, Conjunto Integrativo, Linfonodos Examinados Positivos, Contagem de Mutações, Índice Prognóstico de Nottingham, Código Oncotree (sistema de codificação para classificar tipos de câncer), Sobrevivência Geral (Meses), Tamanho do Tumor, Estágio do Tumor, Tipo de Câncer Detalhado.

Em seguida, valores categóricos foram transformados em valores numéricos para a seleção de características por importância de cada característica com o intuito de melhorar o desempenho e simplificar o modelo, que reduz tempo de treinamento e aumentar sua interoperabilidade. As características selecionadas foram: Idade no diagnóstico, subtipo histológico específico de tumor, Índice Prognóstico de Nottingham (ferramenta prognóstica usada para avaliar o desfecho clínico e o risco de recorrência em pacientes com câncer de mama), Código Oncotre, meses entre o diagnóstico (ou começo do tratamento) até a morte do paciente.

A partir disso, o próximo passo foi tratar do balanceamento das classes, ou seja, garantir que não haja muitos registros de uma classe enquanto haja poucos de outra. Adotando o método SMOTE (Synthetic Minority Over-Sampling Technique, ou seja, Técnica de Superamostragem Sintética de Minorias), onde novas amostras sintéticas são geradas a partir da combinação de amostras minoritárias existentes – porém, para a geração dessas amostras sintéticas são necessárias um número mínimo de classes originais, logo, as classes Angiossarcoma de Mama e Câncer de Mama Metaplástico foram unidas em uma só por terem poucas amostras –, o balanceamento bem como a divisão de dados de treino e teste para: dados brutos e dados normalizados. Na figura 8 é exibida a porcentagem de distribuição de cada classe após esse processo.

Figura 8 – Distribuição das Classes Após Balanceamento



Fonte: Elaborado pelo autor

3.3 K-ésimo Vizinho Mais Próximo

Para a aplicação do método KNN já explicado no capítulo 2, foi necessário determinar um número K apropriado para a classificação. A seleção de tal número foi feita por meio da

chamada Cross-validation (Validação em cruz), onde o modelo ao invés de ser treinado e testado uma única vez têm-se seu conjunto de dados dividido em partes chamadas de folds (repartições), realizando o treinamento e teste K vezes (neste caso, de 1 a 50 para K e usando 10 folds). Apesar de exigir um custo computacional a mais, a determinação de um K é feita a partir de uma pontuação atribuída para execução e no fim do processo, o K é determinado pelo índice do vetor com maior pontuação.

Com isso, o método KNN foi executado para dados brutos e normalizados, os resultados serão exibidos no próximo capítulo deste trabalho.

3.4 Árvore de Decisão e Floresta Randômica

A implementação dos métodos de classificação de árvore de decisão e floresta randômica foram realizadas por meio da biblioteca python sklearn. Novamente, as devidas calibragens como profundidade máxima da árvore, mínimo de amostras para dividir um nó em 40 e mínimo de amostras em um nó de folha em 20, foram devidamente ajustadas de acordo com o tamanho da base de dados e as features selecionadas anteriormente.

Assim como o método KNN, estes métodos foram executados com dados brutos e dados normalizados e os resultados serão exibidos no capítulo posterior.

3.5 Aprendizado Zero-shot

Para a implementação do modelo ZSL foi necessária uma etapa de pré-processamento extra onde houve a conversão da base de dados em uma vertente textual. Esta tarefa foi efetuada, onde cada registro de cada coluna foi traduzido em um texto, que posteriormente foram concatenadas de modo a transformar uma linha inteira em uma frase textual (em inglês). Também, houve a definição textual de cada rótulo (ou classe) que foram possibilidade chamados de rótulos candidatos.

Após essa preparação dos dados, o modelo escolhido para a aplicação do algoritmo foi o SetFit da HuggingFace com um total de 10 épocas. Ao preparar o ambiente, a classificação é feita da seguinte forma: o modelo transforma os textos de cada registro em vetores de alta dimensão que representam o significado semântico do texto chamados *embbedings*, então o SetFit utiliza descrições das classes para guiar o modelo no processo de aprendizagem, possibilitando que aprenda características que distinguem as classes uma da outra e, por fim,

classifica dados nunca antes vistos, baseando-se nas descrições das classes e *embbedings* previamente vistos.

3.6 Considerações finais

Em suma, o processo desenvolvido gerou a implementação de forma a adquirir os dados provenientes da saúde, adequá-los por meio de pré-processamento e limpeza para então submeter estes dados a uma análise. Tal análise sugeriu a aplicação de técnicas para balanceamento e seleção de características principais, as quais foram aplicadas para por fim, realizar-se a classificação dos dados com os métodos convencionais sugeridos e o Zero-Shot-Learning.

4 Avaliação Experimental

Neste capítulo são apresentados os resultados obtidos por meio da implementação apresentada no capítulo anterior bem como as comparações entre as métricas.

4.1 Método de Avaliação

Os modelos previamente citados foram avaliados a partir de quatro métricas de avaliação de modelos de IA que são:

- Acurácia - métrica mais básica de avaliação de um modelo. Trata-se da proporção do número de predições feita de forma correta pelo número total de predições.
- Precisão - representa a quantidade de predições corretas em relação ao total de predições, ou seja, o total de verdadeiros positivos dividido pela soma de verdadeiros positivos e falsos positivos.
- Recall (ou sensibilidade) - semelhante a precisão, é o resultado da proporção de predições positivas corretas em relação ao total de exemplos que realmente são positivos no conjunto, ou seja, verdadeiros positivos dividido por verdadeiros positivos somados com falsos negativos.
- F1 Score - é uma média harmônica entre a precisão e o recall.

Além das métricas, foram apresentadas as tabelas contendo os resultados e matrizes de confusão para a melhor compreensão do número de verdadeiros positivos, falsos positivos, verdadeiros negativos e falsos negativos.

Após a apresentação dos resultados alcançados, uma discussão é efetuada ao final do capítulo, em que são comparadas as diferenças e analisa-se se o objetivo original foi alcançado.

4.2 Resultados

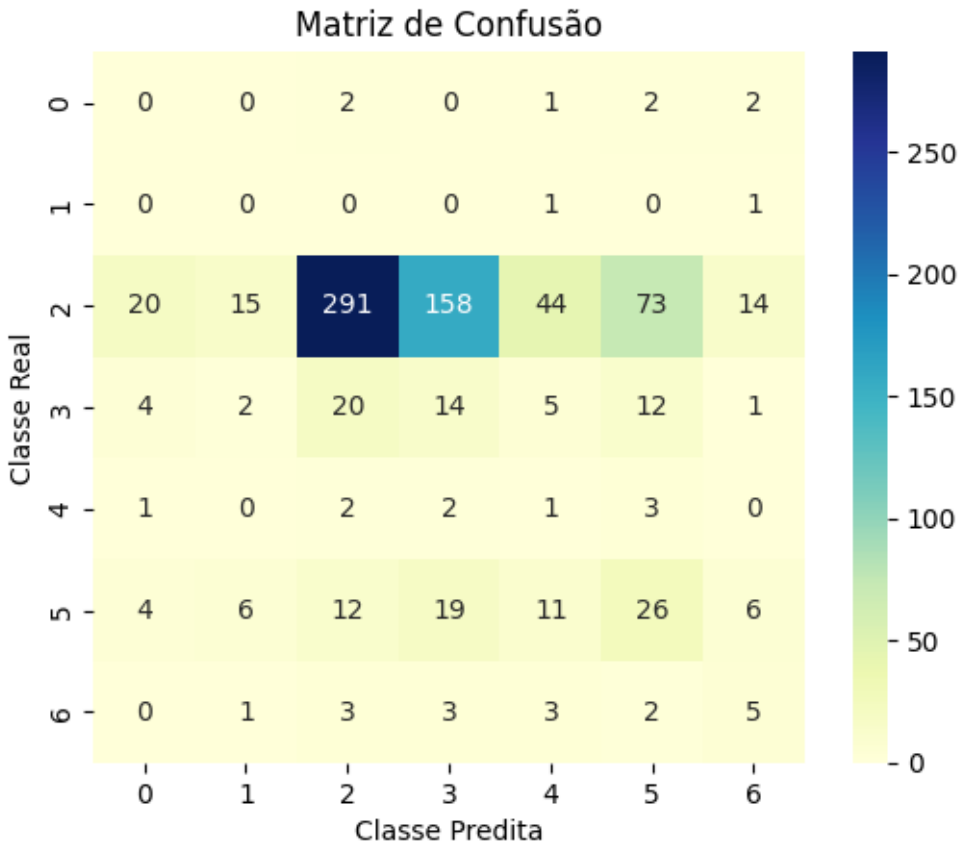
Os primeiros resultados a serem avaliados foram os relativos ao modelo KNN. Foram exibidos na respectiva ordem: a tabela com os resultados gerais do KNN, a matriz de confusão para dados brutos e em seguida para dados normalizados. Na tabela 1 são demonstradas as métricas, explicadas anteriormente, nas quais foram constatadas um desempenho ruim quando foram utilizados dados brutos, porém para os dados normalizados foram obtidos resultados satisfatórios, isto é, por tratar-se de dados vinculados a uma área crítica como a saúde, valores acima de 0.9 são aceitáveis. Abaixo, é exibido na figura 9, pela matriz de confusão quais os números que geraram os resultados da tabela 1, onde o eixo horizontal apresenta quais classes foram previstas e o vertical os rótulos verdadeiros, ou seja, uma célula que represente a intersecção de classe prevista e classe real é uma previsão correta e vice-e-versa. Dessa forma, os dados apresentados na tabela 1 correspondem a matriz de confusão após análise posterior, pois é possível observar que para dados brutos existem muitas predições que não correspondem as suas respectivas classes reais. Da mesma forma, para dados normalizados a diagonal apresenta mais resultados, o que seguindo a lógica explicada anteriormente, corresponde a mais predições feitas de forma correta.

Tabela 1 – Resultados Obtidos do Modelo KNN

Resultados KNN				
	Acurácia	**Precisão**	**Recall**	**F1 - Score**
Dados Brutos	0.42	0.72	0.43	0.52
Dados Normalizados	0.91	0.93	0.92	0.92

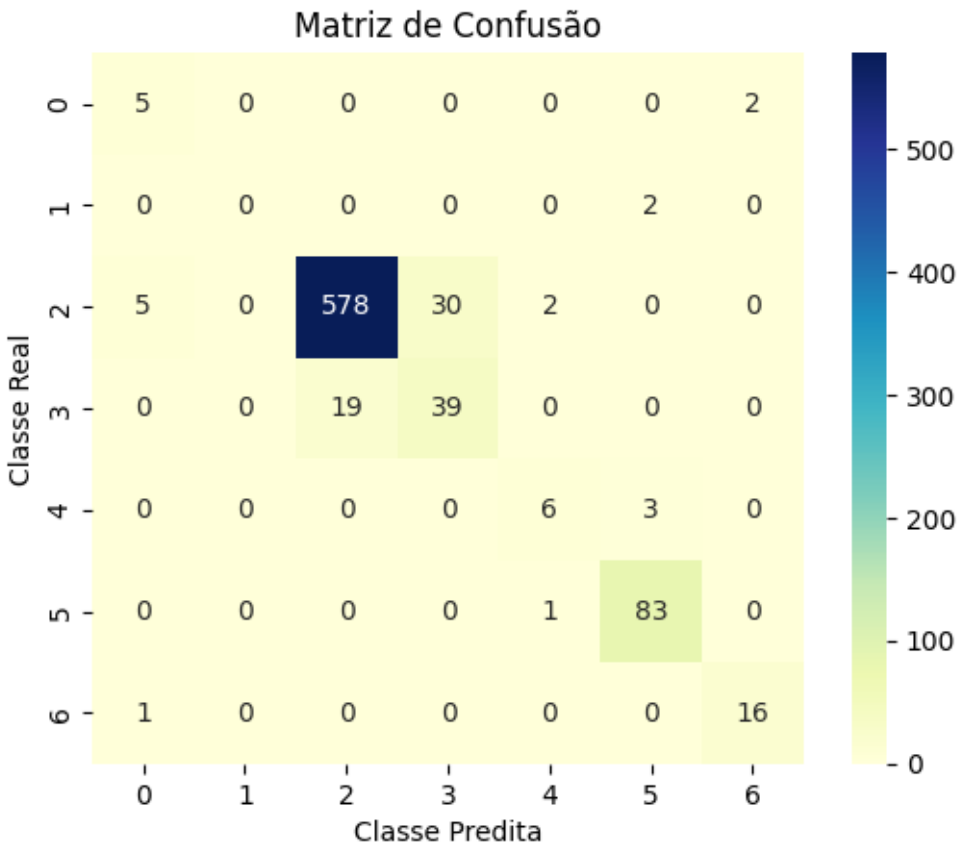
Fonte – Elaborado pelo autor

Figura 8 – Matriz de Confusão para Dados Brutos do Modelo KNN



Fonte – Elaborado pelo autor

Figura 9 – Matriz de Confusão do Modelo KNN para Dados Normalizados



Fonte – Elaborado pelo autor

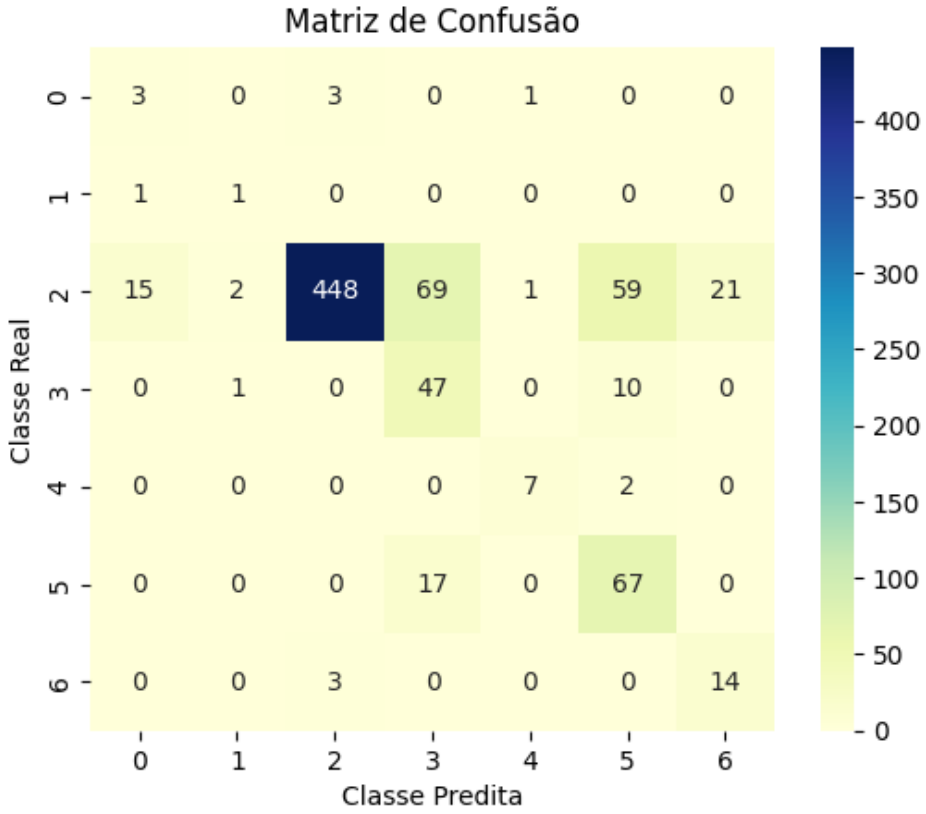
Em seguida, segue os resultados utilizando o modelo de árvore de decisão para dados brutos e dados normalizados. A partir dos resultados obtidos demonstrados na tabela 2, é notável tanto uma melhora quanto uma queda de performance em relação ao modelo anterior. Para dados brutos o modelo de árvore de decisão originou resultados melhores do que o KNN, porém resultados inferiores para dados normalizados. Todavia, o modelo de árvore de decisão não ultrapassou o limiar aceitável de 0.9 para dados críticos, o que também é refletido nas matrizes de confusão das figuras 11 e 12 consequentemente.

Tabela 2 – Resultados do Modelo Árvore de Decisão

Resultados árvore de decisão				
	Acurácia	**Precisão**	**Recall**	**F1 - Score**
Dados Brutos	0.74	0.86	0.74	0.77
Dados Normalizados	0.77	0.86	0.77	0.80

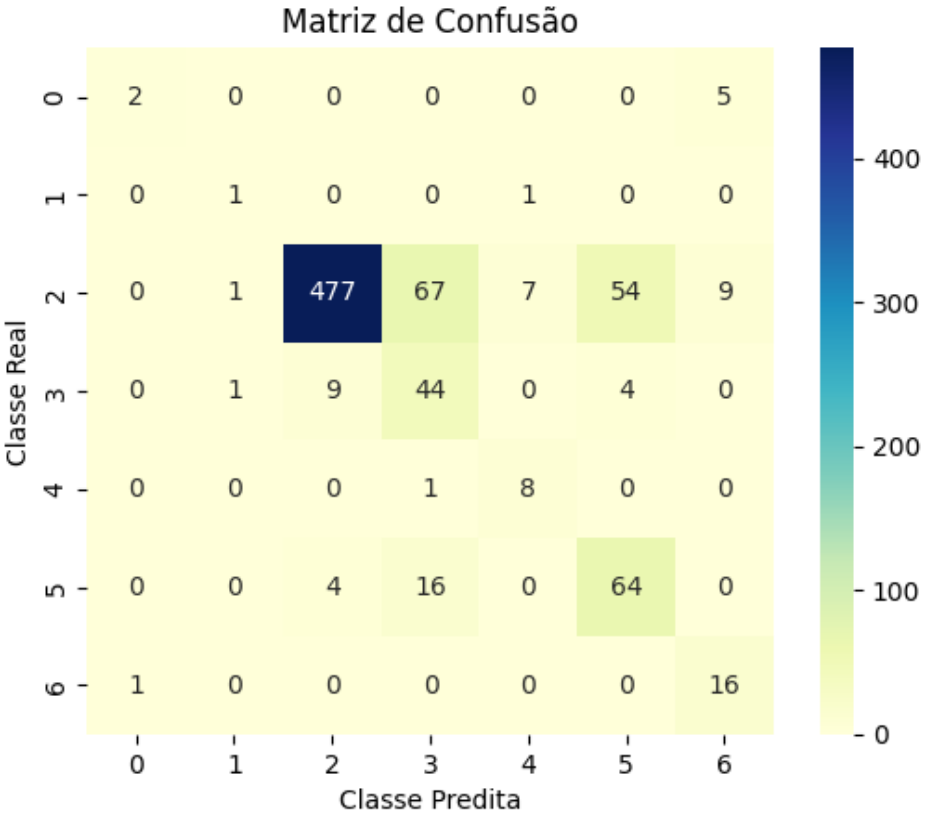
Fonte – Elaborado pelo autor

Figura 10 – Matriz de Confusão do Modelo Árvore de Decisão com Dados Brutos



Fonte – Elaborado pelo autor

Figura 11 – Matriz de Confusão do Modelo Árvore de Decisão com Dados Normalizados



Fonte – Elaborado pelo autor

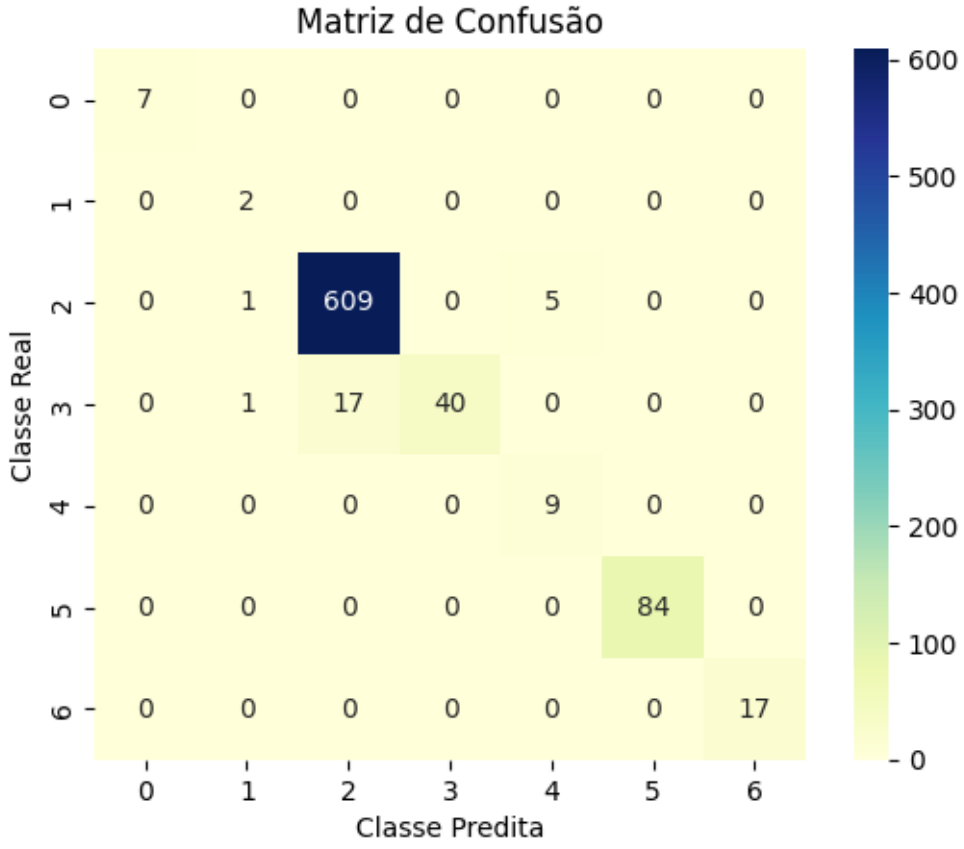
Em seguida, segue os resultados que utilizam o modelo de floresta aleatória para dados brutos e dados normalizados. É exibido na tabela 3 os resultados do modelo de floresta aleatória, os quais superaram os dois modelos anteriores tanto para dados brutos quanto dados normalizados não só ultrapassando o corte de 0.9 como se excedendo até 0.95, novamente sendo refletidos nas matrizes de confusão das figuras 13 e 14 com apenas alguns números discernindo da diagonal principal da matriz.

Tabela 3 – Resultados Floresta Aleatória

Resultados floresta aleatória				
	Acurácia	**Precisão**	**Recall**	**F1 - Score**
Dados Brutos	0.96	0.97	0.97	0.96
Dados Normalizados	0.98	0.96	0.97	0.98

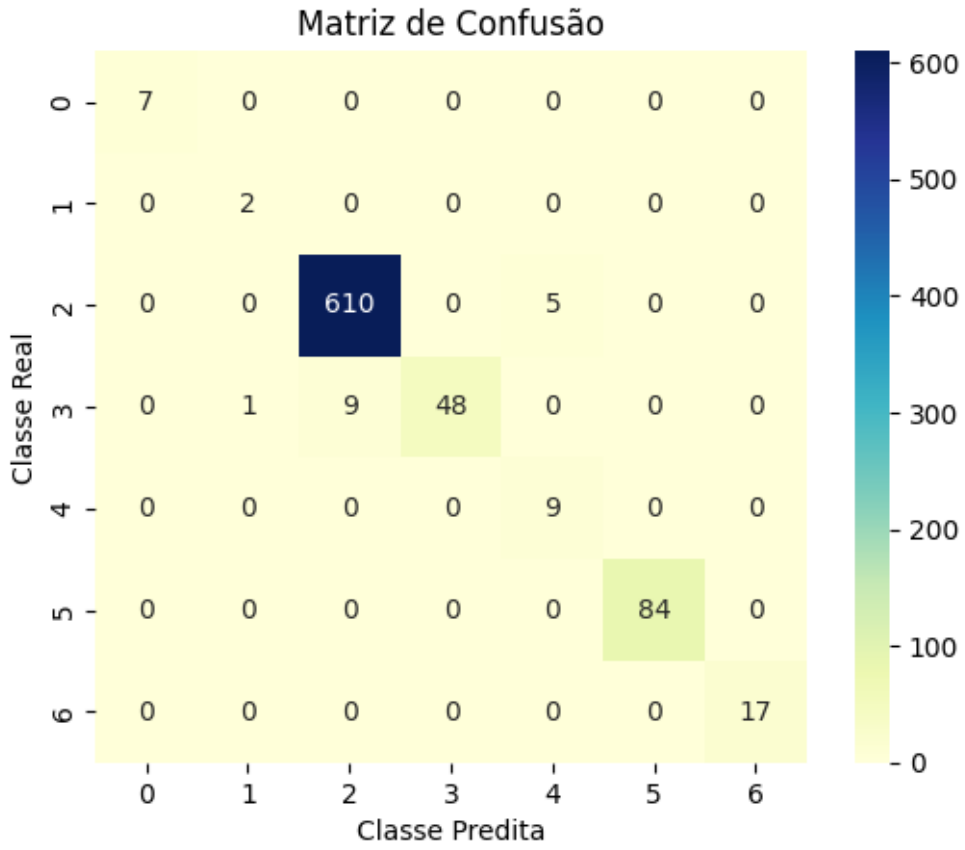
Fonte: Elaborado pelo autor

Figura 12 – Matriz de Confusão do Modelo Floresta Aleatória com Dados Brutos



Fonte – Elaborado pelo autor

Figura 13 – Matriz de Confusão do Modelo Floresta Aleatória com Dados Normalizados



Fonte – Elaborado pelo autor

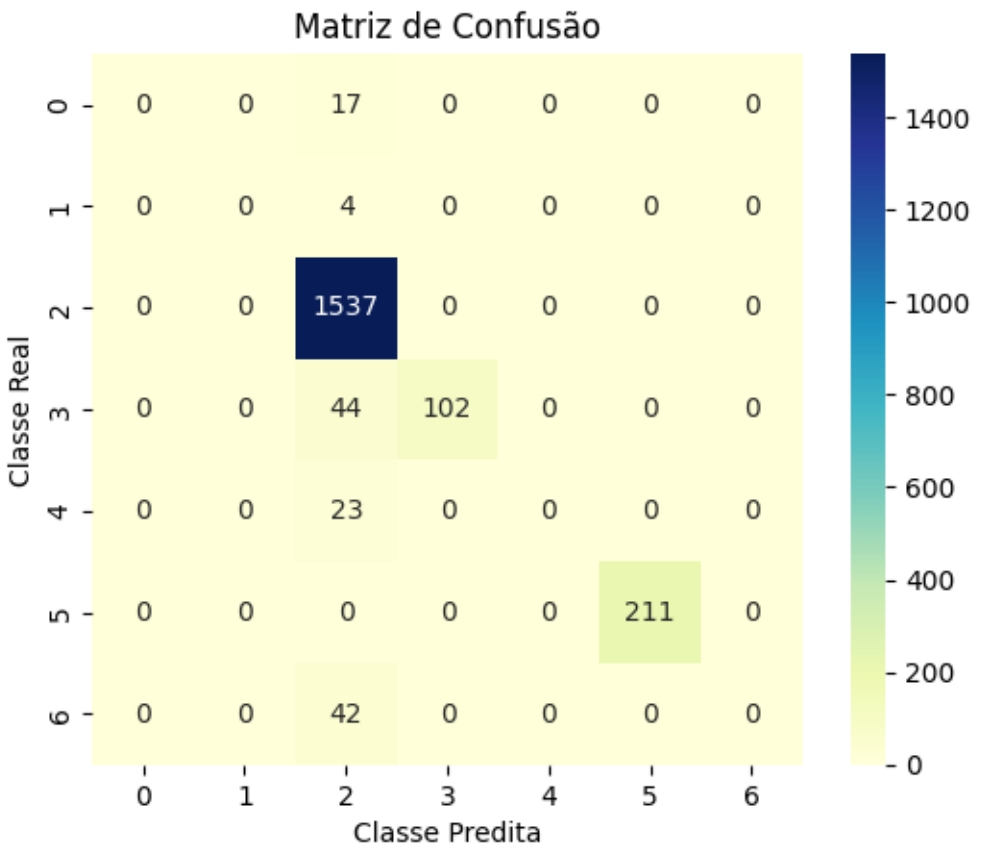
Por fim, o modelo ZSL foi aplicado para dados brutos, tendo em vista a conversão deles para uma forma textual e a técnica de normalização usada nos outros modelos, a qual transforma os dados em vetores numéricos. É exibido na tabela 4 os resultados obtidos, todos acima do limiar de 0.9, o que reflete em uma performance acima dos métodos KNN e árvore de decisão para esse tipo de dado.

Tabela 4 – Resultados ZSL

Resultados ZSL				
	Acurácia	**Precisão**	**Recall**	**F1 - Score**
Dados Brutos	0.93	0.94	0.93	0.91

Fonte – Elaborado pelo autor

Figura 14 – Matriz de Confusão do Modelo ZSL



Fonte – Elaborado pelo autor

4.3 Discussões Sobre os Resultados Obtidos

É notável pelos resultados atingidos e demonstrados pelos gráficos e tabelas apresentados que o modelo ZSL atingiu o limiar de 0.9, o qual não foi atingido em todos os aspectos pelos modelos KNN e árvore de decisão. Apesar de não ser o melhor resultado obtido, demonstrou-se ser uma opção viável de classificação de dados sensíveis mesmo sendo aplicado apenas para dados brutos enquanto os outros modelos tiveram acesso a dados numéricos normalizados.

Dessa forma, a proposta deste trabalho foi alcançada. Os resultados obtidos destacam o potencial do ZSL, que originalmente se destacou na classificação de imagens, em expandir seu uso para dados tabulares, demonstrando-se uma abordagem promissora. A partir disso, o ZSL demonstra potencial na esfera da saúde, onde os dados como exibido por Workman et al. (2023) no capítulo anterior, podem ser muito mais descritivos do que numéricos.

5 Conclusão

A evolução dentro do contexto de IA nos últimos anos resulta em frutos em diversas áreas da ciência. A interdisciplinaridade entre a tecnologia e a esfera da saúde mantém-se forte devido ao aprimoramento de técnicas e aparatos que contribuem de forma significativa para a manutenção e avanço da saúde humana. Dentro desse espectro, quando se trata da análise de dados, a presença de metodologias que envolvam IA se faz inerente devido ao grande volume de dados produzidos diariamente, dados estes que não são coletados por especialistas de dados, resultando em padrões diferentes, textuais, dados faltantes, dentre outros.

Com isso em mente, o trabalho vigente propôs uma abordagem na qual o modelo ZSL provou-se eficaz quando lidando com dados tabulares em comparação com modelos convencionais de classificação. Apesar de em termos de aplicabilidade os modelos clássicos ainda serem preferíveis em um cenário onde há um grande volume de rótulos disponíveis, o modelo ZSL demonstrou ser uma alternativa promissora para situações em que há escassez de dados rotulados, como em cenários de diagnósticos médicos.

Portanto, conclui-se que o ZSL complementa a abordagem dos modelos tradicionais ao oferecer uma solução robusta e flexível para problemas de predição na área da saúde, possibilitando sua utilização em ambientes onde haja limitações de dados e a necessidade de generalização representam desafios significativos. Como será sugerido na seção posterior, essa metodologia pode vir a ser expandida em futuros estudos, integrando outros tipos de dados, além de explorar arquiteturas mais avançadas de ZSL com o intuito de melhorar a capacidade de predição em cenários de grande complexidade.

5.1 Contribuições Científicas

A utilização do modelo ZSL para classificação em diversas áreas da literatura científica cresce como um tema cada vez mais recorrente. A oportunidade de realizar o uso de uma IA para classificar classes nunca antes vistas gera uma gama de possibilidades a serem exploradas

em esferas do conhecimento, dentre elas, a saúde como demonstrado pelos trabalhos correlatos apresentados anteriormente. Assim sendo, a contribuição científica deste trabalho trata-se da comparação de um método emergente como o ZSL com métodos já consolidados de classificação de dados ao utilizar dados vinculados a saúde, onde mostra-se eficaz com dados tabulares e promissor para situações em que há escassez de dados rotulados.

Na tabela 5 são apresentadas as comparações das características dos trabalhos correlatos apresentados no capítulo 2 e este trabalho.

Tabela 5 – Comparativo Entre os Trabalhos Correlatos e o Trabalho Desenvolvido

	(FARHAD et al., 2023)	**(WORKMAN et al., 2023)**	**(BHATE et al., 2023)**	**(KAUFMANN et al., 2023).**	**Este trabalho**
Utilização de ZSL juntamente com dados vinculados a saúde	x	x	x	x	x
Cenário com classes desconhecidas, favorecendo o uso de ZSL	x	x	x	x	x
Utilização de ZSL juntamente com outras técnicas de classificação (conhecidas ou não)	x		x		x
Base de dados já dispostas na literatura anteriormente					x
Classificação de dados tabulares com ZSL					x
Comparação com técnicas amplamente conhecidas de classificação	x		x		x
Métricas maiores ou iguais a 0.9	x	x		x	x

Fonte: Elaborado pelo autor

5.2 Trabalhos Futuros

Como sugerido anteriormente, há pontos que podem ser abordados em trabalhos futuros como:

1. Comparação de métodos convencionais juntamente com ZSL e Few-Shot-Learning
2. Abordagem de Embbedings – vetores criados por modelos de aprendizado de máquina – mais sofisticados como o GPT e BERT com o intuito de aprimorar a representação das classes não vistas e aumentar a capacidade do modelo.

3. Expandir o modelo para outros dados clínicos, tabulares ou não, de forma a melhorar a auxiliar a formação de diagnósticos mais precisos por profissionais da saúde em meio a falta de informações.

REFERÊNCIAS

YANG, Jin et al. Brief introduction of medical database and data mining technology in big data era. **Journal of Evidence-Based Medicine**, v. 13, n. 1, p. 57-69, 2020.

TRIFIRÒ, Gianluca; SULTANA, Janet; BATE, Andrew. From big data to smart data for pharmacovigilance: the role of healthcare databases and other emerging sources. **Drug safety**, v. 41, p. 143-149, 2018.

KAMM, Simon; JAZDI, Nasser; WEYRICH, Michael. Knowledge discovery in heterogeneous and unstructured data of industry 4.0 systems: challenges and approaches. **Procedia CIRP**, v. 104, p. 975-980, 2021.

LATTAR, Hafsa; SALEM, Aicha Ben; GHEZALA, Henda Hajjami Ben. Does data cleaning improve heart disease prediction?. **Procedia Computer Science**, v. 176, p. 1131-1140, 2020.

GUPTA, Rohan et al. New era of artificial intelligence and machine learning-based detection, diagnosis, and therapeutics in Parkinson's disease. **Ageing research reviews**, p. 102013, 2023.

OLAWADE, David B. et al. Using artificial intelligence to improve public health: a narrative review. **Frontiers in Public Health**, v. 11, p. 1196397, 2023.

SHAMSHIRBAND, Shahab et al. A review on deep learning approaches in healthcare systems: Taxonomies, challenges, and open issues. **Journal of Biomedical Informatics**, v. 113, p. 103627, 2021.

TAN, Ting Fang et al. Artificial intelligence and digital health in global eye health: opportunities and challenges. **The Lancet Global Health**, v. 11, n. 9, p. e1432-e1443, 2023.

World Health Organization. (2019). WHO guideline: recommendations on digital interventions for health system strengthening. World Health Organization. https://iris.who.int/handle/10665/311941 (acessado em 05/06/2024).

NGIAM, Kee Yuan; KHOR, Wei. Big data and machine learning algorithms for health-care delivery. **The Lancet Oncology**, v. 20, n. 5, p. e262-e273, 2019.

POURPANAH, Farhad et al. A review of generalized zero-shot learning methods. **IEEE transactions on pattern analysis and machine intelligence**, v. 45, n. 4, p. 4051-4070, 2022.

ADADI, Amina. A survey on data-efficient algorithms in big data era. **Journal of Big Data**, v. 8, n. 1, p. 24, 2021.

KITCHIN, Rob. **The data revolution: Big data, open data, data infrastructures and their consequences**. Sage, 2014.

AGRAWAL, Ankit; CHOUDHARY, Alok. Health services data: big data analytics for deriving predictive healthcare insights. **Health services evaluation**, p. 3-18, 2019.

AL MAYAHI, Salma; AL-BADI, Ali; TARHINI, Ali. Exploring the potential benefits of big data analytics in providing smart healthcare. In: **Emerging Technologies in Computing: First International Conference, iCETiC 2018, London, UK, August 23–24, 2018, Proceedings 1**. Springer International Publishing, 2018. p. 247-258.

FANG, Hua et al. A survey of big data research. **IEEE network**, v. 29, n. 5, p. 6-9, 2015.

GANDOMI, Amir; HAIDER, Murtaza. Beyond the hype: Big data concepts, methods, and analytics. **International journal of information management**, v. 35, n. 2, p. 137-144, 2015.

OLSZAK, Celina M.; MACH-KRÓL, Maria. A conceptual framework for assessing an organization's readiness to adopt big data. **Sustainability**, v. 10, n. 10, p. 3734, 2018.

RISTEVSKI, Blagoj; CHEN, Ming. Big data analytics in medicine and healthcare. **Journal of integrative bioinformatics**, v. 15, n. 3, p. 20170030, 2018.

LEE, Choong Ho; YOON, Hyung-Jin. Medical big data: promise and challenges. **Kidney research and clinical practice**, v. 36, n. 1, p. 3, 2017.

OSSIO, Raul et al. Melanoma: a global perspective. **Nature Reviews Cancer**, v. 17, n. 7, p. 393-394, 2017.

RUMSFELD, John S.; JOYNT, Karen E.; MADDOX, Thomas M. Big data analytics to improve cardiovascular care: promise and challenges. **Nature Reviews Cardiology**, v. 13, n. 6, p. 350-359, 2016.

DRUCKER, Johanna. Humanities approaches to graphical display. **Digital Humanities Quarterly**, v. 5, n. 1, 2011.

BATKO, Kornelia; ŚLĘZAK, Andrzej. The use of Big Data Analytics in healthcare. **Journal of big Data**, v. 9, n. 1, p. 3, 2022.

FRAWLEY, William J.; PIATETSKY-SHAPIRO, Gregory; MATHEUS, Christopher J. Knowledge discovery in databases: An overview. **AI magazine**, v. 13, n. 3, p. 57-57, 1992.

FAYYAD, Usama; PIATETSKY-SHAPIRO, Gregory; SMYTH, Padhraic. From data mining to knowledge discovery in databases. **AI magazine**, v. 17, n. 3, p. 37-37, 1996.

GARCÍA, Salvador et al. Big data preprocessing: methods and prospects. **Big data analytics**, v. 1, p. 1-22, 2016.

CHU, Xu et al. Data cleaning: Overview and emerging challenges. In: **Proceedings of the 2016 international conference on management of data**. 2016. p. 2201-2206.

RAHM, Erhard; DO, Hong Hai. Data cleaning: Problems and current approaches. **IEEE Data Eng. Bull.**, v. 23, n. 4, p. 3-13, 2000.

MIKUT, Ralf; REISCHL, Markus. Data mining tools. **Wiley interdisciplinary reviews: data mining and knowledge discovery**, v. 1, n. 5, p. 431-443, 2011.

CHEN, Kaile et al. Process mining and data mining applications in the domain of chronic diseases: A systematic review. **Artificial Intelligence in Medicine**, p. 102645, 2023.

MATIGNON, Laëtitia; LAURENT, Guillaume J.; LE FORT-PIAT, Nadine. Hysteretic q-learning: an algorithm for decentralized reinforcement learning in cooperative multi-agent teams. In: **2007 IEEE/RSJ International Conference on Intelligent Robots and Systems**. IEEE, 2007. p. 64-69.

SZEGEDY C, ZAREMBA W, SUTSKEVER I, BRUNA J, ERHAN D, GOODFELLOW I, FERGUS R. Intriguing properties of neural networks. https://arxiv.org/ abs/1312.6199. 2013.

ZHANG, Tong et al. Hierarchical lifelong learning by sharing representations and integrating hypothesis. **IEEE Transactions on Systems, Man, and Cybernetics: Systems**, v. 51, n. 2, p. 1004-1014, 2019.

ATHEY, Susan. Beyond prediction: Using big data for policy problems. **Science**, v. 355, n. 6324, p. 483-485, 2017.

KRISHNAMOORTHI, Raja et al. [Retracted] A Novel Diabetes Healthcare Disease Prediction Framework Using Machine Learning Techniques. **Journal of healthcare engineering**, v. 2022, n. 1, p. 1684017, 2022.

SCHWALBE, Nina; WAHL, Brian. Artificial intelligence and the future of global health. **The Lancet**, v. 395, n. 10236, p. 1579-1586, 2020.

BOUDREAUX, Edwin D. et al. Applying machine learning approaches to suicide prediction using healthcare data: overview and future directions. **Frontiers in psychiatry**, v. 12, p. 707916, 2021.

BADAWY, Mohammed; RAMADAN, Nagy; HEFNY, Hesham Ahmed. Healthcare predictive analytics using machine learning and deep learning techniques: a survey. **Journal of Electrical Systems and Information Technology**, v. 10, n. 1, p. 40, 2023.

DHALL, Devanshi; KAUR, Ravinder; JUNEJA, Mamta. Machine learning: a review of the algorithms and its applications. **Proceedings of ICRIC 2019: Recent innovations in computing**, p. 47-63, 2020.

YANG, Feng-Jen. An extended idea about decision trees. In: **2019 international conference on computational science and computational intelligence (CSCI)**. IEEE, 2019. p. 349-354.

EESA, Adel Sabry; ORMAN, Zeynep; BRIFCANI, Adnan Mohsin Abdulazeez. A novel feature-selection approach based on the cuttlefish optimization algorithm for intrusion detection systems. **Expert systems with applications**, v. 42, n. 5, p. 2670-2679, 2015.

GUPTA, Monica; PANDYA, Sohil D. A comparative study on supervised machine learning algorithm. **International Journal for Research in Applied Science and Engineering Technology (IJRASET)**, v. 10, n. 1, p. 1023-1028, 2022.

BAKYARANI, E. Sweety et al. A survey of machine learning algorithms in health care. **Int J Sci Technol Res**, v. 8, n. 11, p. 223, 2019.

SRIVASTAVA, Aditya; SAINI, Sonia; GUPTA, Deepa. Comparison of various machine learning techniques and its uses in different fields. In: **2019 3rd International conference on electronics, communication and aerospace technology (ICECA)**. IEEE, 2019. p. 81-86.

HO, Tin Kam. The random subspace method for constructing decision forests. **IEEE transactions on pattern analysis and machine intelligence**, v. 20, n. 8, p. 832-844, 1998.

HOFMANN, Markus; KLINKENBERG, Ralf (Ed.). **RapidMiner: Data mining use cases and business analytics applications**. CRC Press, 2016.

CHOW, C. K. C. N.; LIU, Cong. Approximating discrete probability distributions with dependence trees. **IEEE transactions on Information Theory**, v. 14, n. 3, p. 462-467, 1968.

ROMERA-PAREDES, Bernardino; TORR, Philip. An embarrassingly simple approach to zero-shot learning. In: **International conference on machine learning**. PMLR, 2015. p. 2152-2161.

BHATE, Neel Jitesh et al. Zero-shot learning with minimum instruction to extract social determinants and family history from clinical notes using GPT model. In: **2023 IEEE International Conference on Big Data (BigData)**. IEEE, 2023. p. 1476-1480.

DONG, Pengzhi et al. Identification of key genes and pathways in triple-negative breast cancer by integrated bioinformatics analysis. **BioMed Research International**, v. 2018, n. 1, p. 2760918, 2018.

BLOCKHUYS, Stéphanie; BRADY, Donita C.; WITTUNG-STAFSHEDE, Pernilla. Evaluation of copper chaperone ATOX1 as prognostic biomarker in breast cancer. **Breast Cancer**, v. 27, p. 505-509, 2020.

KAUFMANN, Basil et al. Validation of a Zero-Shot Learning Natural Language Processing Tool for Data Abstraction from Unstructured Healthcare Data. **arXiv preprint arXiv:2308.00107**, 2023.