UNIVERSIDADE FEDERAL DO MARANHÃO

Programa de Pós-Graduação em Ciência da Computação

Lucelia Lima Souza

Descoberta de Conhecimento nas Bases de Dados da Pandemia da COVID-19 e de Indicadores Socioeconômicos e Ambientais

São Luís
2023

Lucelia Lima Souza

Descoberta de Conhecimento nas Bases de Dados da Pandemia da COVID-19 e de Indicadores Socioeconômicos e Ambientais

Dissertação apresentada como requisito parcial para obtenção do título de Mestre em Ciência da Computação, ao Programa de Pós-Graduação em Ciência da Computação, da Universidade Federal do Maranhão.

Programa de Pós-Graduação em Ciência da Computação

Universidade Federal do Maranhão

Orientador: Prof. Dr. Tiago Bonini Borchartt

São Luís - MA

2023

Ficha gerada por meio do SIGAA/Biblioteca com dados fornecidos pelo(a) autor(a).
Diretoria Integrada de Bibliotecas/UFMA

Lima Souza, Lucelia.
Descoberta de conhecimento nas bases de dados da pandemia da COVID-19 e de indicadores socioeconômicos e ambientais / Lucelia Lima Souza. - 2023.
83 f.

Orientador(a): Prof. Dr. Tiago Bonini Borchartt.
Dissertação (Mestrado) - Programa de Pós-graduação em Ciência da Computação/ccet, Universidade Federal do Maranhão, Videoconferência, 2023.

1. COVID-19. 2. Dados Ambientais. 3. Data Mining. 4. Descoberta de Conhecimento. 5. Indicadores socioeconômicos. I. Bonini Borchartt, Prof. Dr. Tiago. II. Título.

Lucelia Lima Souza

Descoberta de Conhecimento nas Bases de Dados da Pandemia da COVID-19 e de Indicadores Socioeconômicos e Ambientais

Dissertação apresentada como requisito parcial para obtenção do título de Mestre em Ciência da Computação, ao Programa de Pós-Graduação em Ciência da Computação, da Universidade Federal do Maranhão.

Trabalho aprovado em 24 de Abril de 2023

Prof. Dr. Tiago Bonini Borchartt
Orientador
Universidade Federal do Maranhão

Prof. Dr. Luciano Reis Coutinho
Universidade Federal do Maranhão

Prof. Dr. Sérgio Teixeira de Carvalho
Universidade Federal de Goiás

São Luís - MA
2023

A Deus, pela vida e oportunidade, e a toda minha família e amigos.

Agradecimentos

Agradeço a Deus em primeiro lugar pela vida, pelos desafios e oportunidades de ter conseguido chegar até aqui, para meu crescimento profissional e pessoal, por estar sempre comigo e nunca me abandonar, em todos momentos me deu forças pra continuar, obrigada meu Deus! Toda Glória seja a ELE.

Aos meus pais, minha mãe Iêda Lima Souza e meu pai Raimundo Nonato Souza *(In memoriam)*, que sempre me deram forças para nunca desistir. Serei sempre grata, pois o ano de 2022, foi muito difícil, por enfrentar a perda do meu pai.

Aos meus irmãos e irmãs, pelo apoio incondicional e por acreditarem nos meus sonhos.

À minha amiga Maria Alice, por ouvir meus lamentos de cansaços, choros e sempre me dar força pra nunca desistir.

Aos colegas de mestrado, pela amizade, em especial ao colega "Luís Eduardo Laurindo"que juntos compartilhamos nossas lutas.

Ao meu orientador o Professor Doutor Tiago Bonini pela paciência e compreensão, em toda a etapa desse processo de orientação, pelos ensinamentos e colaboração.

A todos que ajudaram diretamente e indiretamente por essa conquista, com todo apoio.

"O temor do Senhor é o princípio da sabedoria,
e o conhecimento do Santo a prudência."

(Provérbios 9.10)

Resumo

A pandemia da COVID-19 desencadeou uma crise global de saúde pública e exigiu a análise de dados em larga escala para entender melhor sua disseminação e impacto na sociedade. Neste contexto, a "Descoberta de Conhecimento em Bases de Dados" *(Knowledge Discovery in Databases, KDD)* é uma ferramenta útil, pois apresenta uma metodologia bem definida, com etapas validadas em diferentes aplicações. O presente trabalho objetiva descobertas de conhecimento dos dados entre a COVID-19 e os Indicadores Socioeconômicos e Ambientais, através do uso das técnicas de Mineração de Dados (MD) - *Data Mining*, classificando novos padrões com método do KDD, visando obter a técnica com o maior percentual de acertos. Para o problema em estudo, o método KDD utilizado é composto pelas etapas de: seleção, pré-processamento, transformação, mineração de dados e avaliação. Obteve-se bons resultados com a aplicação dos métodos de mineração de dados descritiva, que envolvem os modelos de correlação, agrupamento e regra de associação, estas foram as técnicas que mais se destacaram, com capacidades de generalização satisfatórias. Os resultados da descoberta de conhecimento em dados da pandemia da COVID-19 podem contribuir para a formulação de políticas públicas e tomada de decisões informatizadas em saúde pública.

Palavras-chave: COVID-19, Indicadores socioeconômicos, Descoberta de Conhecimento, *Data Mining*, Dados Ambientais.

Abstract

The COVID-19 pandemic has triggered a global public health crisis and required large-scale data analysis to better understand its spread and impact on society. In this context, "Knowledge Discovery in Databases" (KDD) is a useful tool, as it presents a well-defined methodology, with validated steps in different applications. The present work aims at discoveries of knowledge of data between COVID-19 and Socioeconomic and Environmental Indicators, through the use of Data Mining (DM) techniques - Data Mining, classifying new patterns with the KDD method, aiming to obtain the technique with the highest percentage of hits. For the problem under study, the KDD method used is composed of the steps of: selection, pre-processing, transformation, data mining and evaluation. Good results were obtained with the application of descriptive data mining methods, which involve correlation, grouping and association rule models, these were the techniques that stood out the most, with satisfactory generalization capabilities. The results of knowledge discovery in data from the COVID-19 pandemic can contribute to public policy formulation and computerized decision making in public health.

Keywords: COVID-19, *Socioeconomic Indicators, Knowledge Discovery, Data Mining, Environmental Data.*

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

ANNs	*Artificial Neural Networks*
EEA	*European Environmental Agency*
ESPII	*Emergência de Saúde Pública de Importância Internacional*
GPR	*Gaussian Process Regression*
IDH	*Índice de Desenvolvimento Humano*
KDD	*Knowledge Discovery in Databases*
LR	*Regressão Linear Simples*
MD	*Mineração dos Dados*
ML	*Machine Learning*
MLP	*Multi-Layer Perceptrons*
OCDE	*Organização para a Cooperação e Desenvolvimento Econômico*
OECD	*Organisation for Economic Co-operation and Development*
OMS	*Organização Mundial da Saúde*
PIB	*Produto Interno Bruto*
RF	*Random Forests*
RNAs	*Redes Neurais Artificiais*
RSI	*Regulamento Sanitário Internacional*
RSL	*Revisão Sistemática da Literatura*
SHAP	*SHapley Additive exPlanations*
SSLPNN	*Shallow Single-Layer Perceptron Neural Network*
SVM	*Support Vector Machines*
UFMA	*Universidade Federal do Maranhão*
WHO	*World Health Organization*

Sumário

1 Introdução

A COVID-19 é uma doença infecciosa causada pelo coronavírus *SARS-CoV-2*, que foi identificado pela primeira vez em *Wuhan*, na China, no final de 2019. A doença pode causar uma série de sintomas, deste febre, tosse seca, dificuldades respiratórias(dispneia), dores de cabeça e pneumonia, como resultando inicial pode causar insuficiência respiratória progressiva e até a morte (ZHOU et al., 2020). A transmissão do vírus ocorre principalmente por meio do contato com gotículas respiratórias infectadas, seja através do ar ou de superfícies contaminadas.

A pandemia da COVID-19 tem afetado profundamente todos os aspectos da vida, desde a saúde pública até a economia global. Os sistemas de saúde em muitos países foram sobrecarregados, com uma escassez de leitos hospitalares e equipamentos médicos, enquanto a economia global sofreu uma grande desaceleração com o fechamento de empresas e perda de empregos. A Organização para Cooperação e Desenvolvimento Econômico - OCDE alertou sobre o impacto negativo que o Coronavírus causaria ao mundo em relação ao sistema econômico internacional (MAGAZZINO; MELE; MORELLI, 2021). A pandemia continua sendo uma grande preocupação global e a luta contra a doença continua, através da vacinação em massa e medidas de saúde pública.

1.1 Problema e sua importância

Analisar a grande quantidade de dados relacionados à COVID-19 e indicadores socioeconômicos e ambientais pode ser uma tarefa complexa, que requer ferramentas e técnicas de análise adequadas.

Nesta dissertação de mestrado é realizada uma análise da relação entre a pandemia da COVID-19 e os indicadores socioeconômicos e ambientais, análise essa através do processo do *Knowledge Discovery in Databases* - KDD ou Descoberta de Conhecimento em Banco de Dados, que tem como finalidade a extração de conhecimento relevante e exploração desses dados por meio de modelagem, reconhecendo os tipos de padrões existentes.

A Descoberta de Conhecimento em Bases de Dados (KDD) é uma abordagem que combina técnicas de mineração de dados, inteligência artificial e estatística para extrair conhecimento útil a partir de grandes conjuntos de dados, a qual implica a realização de uma sequência de passos: seleção, pré-processamento, transformação, *Data Mining*, interpretação e avaliação dos resultados (GOLDSCHMIDT; PASSOS; BEZERRA, 2015).

A análise da COVID-19 e de indicadores socioeconômicos e ambientais é um

campo de grande interesse para a saúde pública e para o planejamento de políticas sociais e econômicas (PROGRAMME, 2020). A utilização do KDD neste contexto permite a identificação de padrões e relações entre os diferentes fatores envolvidos, contribuindo para uma melhor compreensão da doença e de seus impactos, tais como condições de vida, acesso a serviços de saúde e níveis de renda da população. Além disso, a análise de dados de indicadores ambientais pode fornecer informações sobre a relação entre a qualidade do ar, da água e do solo e a incidência da doença.

Assim, o KDD é uma ferramenta importante para o estudo da COVID-19 e seus impactos sociais e ambientais, bem como para a geração de conhecimento que pode subsidiar a tomada de decisões e a implementação de políticas públicas. A análise de dados pode ajudar a prever surtos de COVID-19, identificar grupos de risco, avaliar a eficácia de medidas de controle e monitorar a disseminação da doença em diferentes regiões. Além disso, o uso do KDD pode contribuir para uma abordagem mais integrada e holística da pandemia, levando em conta não apenas os aspectos médicos, mas também os aspectos sociais, econômicos e ambientais envolvidos. Dessa forma, a aplicação do KDD pode ser uma ferramenta poderosa na luta contra a COVID-19 e na promoção da saúde e bem-estar da população.

A descoberta de conhecimento é definida como o método usado para descobrir padrões interessantes, previamente desconhecidos e potencialmente úteis a partir de uma enorme quantidade de dados (SINGHAL, 2022). KDD é um mecanismo estruturado pelo qual conjuntos de dados massivos e complexos definem padrões reais, novos, úteis e compreensíveis (NASSER; BEHADILI, 2022)(ALI; GHAREB, 2022).

O KDD possui um método desde a coleta de dados até a interpretação do conhecimento. O presente estudo segue esse processo, tendo como objetivo buscar a transformação dos dados armazenados em conhecimento. Nesse contexto, a pesquisa pretende analisar a relação entre os principais indicadores socioeconômicos, alguns indicadores ambientais e os dados alarmantes deixados pela COVID-19.

1.2 Objetivos

O objetivo principal deste trabalho é realizar a descoberta de padrões e relações entre os dados da pandemia da COVID-19 e de indicadores socioeconômicos e ambientais. Para isso, será utilizado o método de processo do KDD.

Os objetivos específicos do trabalho são:

- Realizar uma revisão sistemática da literatura para identificar os indicadores socioeconômicos e ambientais mais relevantes relacionados à saúde pública e as técnicas de Aprendizado de Máquina mais utilizadas para análise de dados da COVID-19;

- Identificar os perfis socioeconômicos que influenciam no número de mortes e casos de COVID-19;
- Construir um *dataset* público que contemple os dados de indicadores socioeconômicos e ambientais e os resultados da pandemia da COVID-19;
- Interpretar e avaliar os padrões e as relações identificadas através do processo do KDD dos dados da COVID-19 e dos indicadores socioeconômicos e ambientais.

1.3 Organização da dissertação

Este trabalho está estruturado de forma a facilitar o entendimento do tema abordado. O conteúdo presente foi dividido nas seguintes seções:

O Capítulo 2 descreve a fundamentação teórica. São abordados conceitos do KDD, COVID-19 e dos Indicadores Socioeconômicos e Ambientais.

O Capítulo 3 apresenta todas as etapas da revisão sistemática da literatura, e seus resultados e discussões.

O Capítulo 4 apresenta os trabalhos relacionados encontrados na literatura e destaca as diferenças deste trabalho em relação aos demais e o gráfico do percurso metodológico para melhor compreensão das etapas percorridas.

O Capítulo 5 apresenta as etapas aplicadas do KDD, descrevendo com detalhes cada etapa realizada, de acordo com o fluxograma da metodologia.

O Capítulo 6 descreve as discussões dos resultados obtidos referente às análises relacionais e aos resultados do KDD.

E por fim, no Capítulo 7 têm-se a conclusão do trabalho e apresentação dos trabalhos futuros.

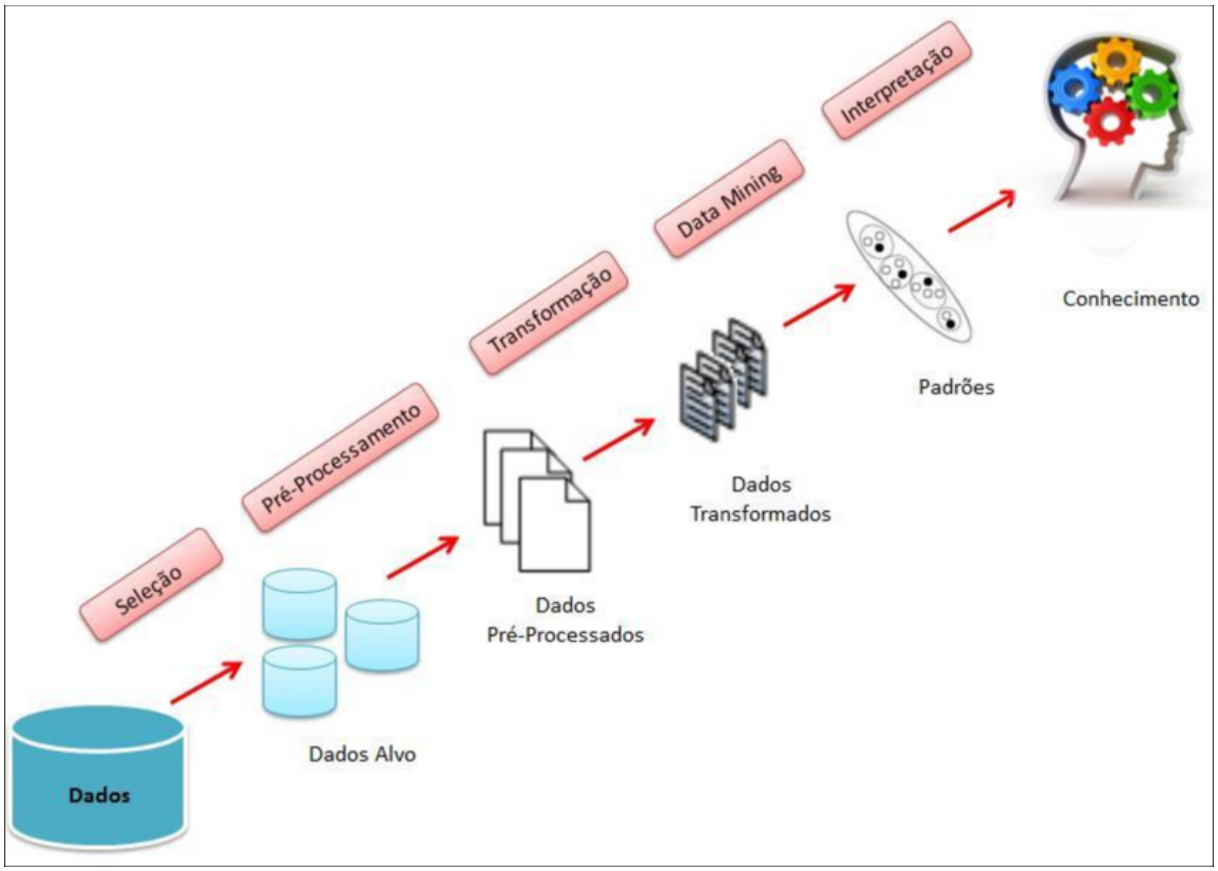
2 Fundamentação Teórica

2.1 *Knowledge Discovery in Databases - KDD*

A Descoberta de Conhecimento em Bancos de Dados (KDD) é uma ferramenta exploratória de análise e modelagem de grandes repositórios de dados, para a descoberta automática de informações. KDD é o processo organizado de identificação de padrões válidos, novos, úteis e compreensíveis de conjuntos de dados grandes e complexos (MAIMON; ROKACH, 2005).

Segundo Fayyad et al. (1996), o KDD é composto de cinco etapas: seleção dos dados, pré-processamento, transformação dos dados, mineração de dados (*Data Mining*); e interpretação e avaliação dos resultados. O processo KDD é interativo e iterativo, envolvendo diferentes etapas, como pode ser observado na Figura 1. A seguir, as etapas do KDD serão descritas.

Figura 1 – Etapas de Processos do KDD



Fonte: Camargo et al. (2016)

2.1.1 Seleção de Dados

Esta primeira etapa tem uma grande importância em relação ao resultado final. Ela consiste em selecionar um conjunto ou um subconjunto de dados, ou seja, as variáveis (atributos) e os registros (instâncias) que farão parte da análise no processo de descoberta. A seleção de dados analisa quais dados são realmente relevantes na base, porque muitas vezes se descobre que nem todos os dados são necessários no processo de mineração, conforme (SUWIRYA et al., 2022).

2.1.2 Pré-processamento de Dados

Após a etapa de seleção, a próxima etapa consiste em realizar a verificação da qualidade dos dados. Neste ponto entram em ação técnicas como o processo de limpeza, correção dos dados da seleção, assim como a remoção de ruídos e de dados duplicados ou discrepantes (*outliers*). De acordo com Maimon e Rokach (2005), esta fase de pré-processamento e limpeza consiste em alcançar a confiabilidade dos dados que passarão para as próximas etapas, podendo ainda ser realizado um aprimoramento neste processo à medida que os dados são analisados.

2.1.3 Transformação de Dados

A etapa de transformação consiste em modelar os atributos que serão úteis para a análise, aplicando técnicas de transformação, como: normalização, padronização, criação de novos atributos, redução, discretização e sintetização dos dados. Essa etapa é onde os dados são transformados ou consolidados de forma apropriada para a mineração, realizando operações de resumo ou agregação conforme (ZHANG; ZHANG; WU, 2004).

2.1.4 Mineração de Dados

Na etapa de Mineração de Dados (MD) são definidas as técnicas e os algoritmos para identificação de padrões nos dados. Fayyad et al. (1996) afirmam que a Mineração de Dados é uma metodologia de resolução de problemas, que encontra uma descrição lógica ou matemática, eventualmente de natureza complexa, de padrões e regularidades em um conjunto de dados.

2.1.5 Interpretação e Avaliação de Resultados

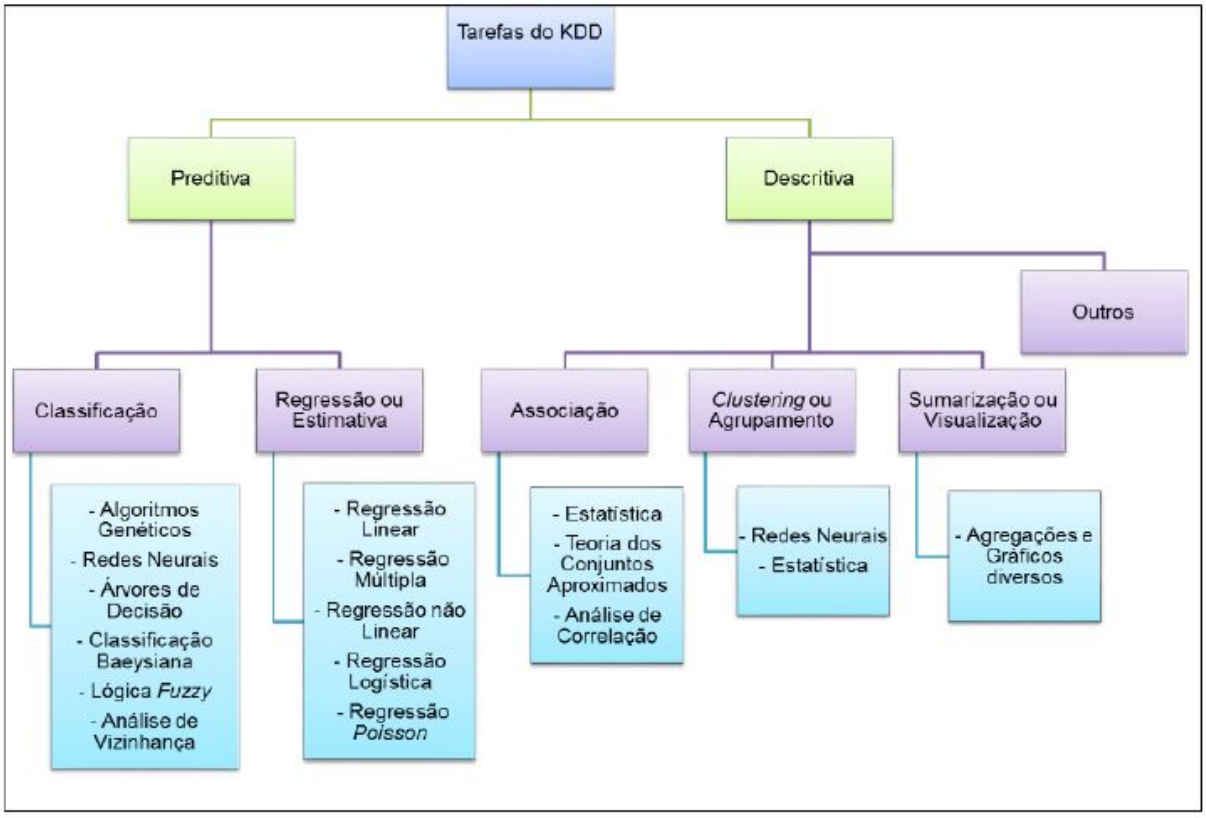
A última etapa do processo do KDD consiste em interpretar os resultados que foram alcançados na etapa de mineração, consolidando o conhecimento descoberto. A avaliação pode ser realizada através da análise de um profissional ou então comparando com os dados que foram inicialmente coletados. Em caso de não satisfação, pode-se retornar às

etapas anteriores. Essa etapa também é conhecida como pós-processamento, envolvendo a visualização, análise e a interpretação do modelo de conhecimento gerado pela etapa de MD (GOLDSCHMIDT; PASSOS; BEZERRA, 2015).

2.2 Tarefas de Processo do KDD

As principais tarefas do KDD têm como base algumas alternativas de métodos de mineração de dados, podendo ser descritivas ou preditivas, na busca de padrões de interesse em uma determinada forma de representação ou um conjunto de tais representações, como: Regras de Classificação ou Regressão, Agrupamento, Associação, Sumarização e assim por diante (FAYYAD et al., 1996), conforme ilustrado na Figura 2.

Figura 2 – Tarefas de KDD e Métodos de Mineração de Dados



Fonte: Ferreira, Rosa e Steiner (2018)

2.2.1 Classificação

A classificação é a tarefa mais estudada no processo KDD, tendo como objetivo encontrar um conhecimento que possa ser empregado para predizer a classe de um determinado registro. Assim sendo, a classificação procura estudar um conjunto de registros históricos (atributos) e elaborar descrições de suas características em cada uma das classes (KAMSU-FOGUEM; RIGAL; MAUGET, 2013).

2.2.2 Regressão

Regressão é utilizada para definição de valor por alguma variável contínua. Esta técnica tem como objetivo aprender uma função que mapeia um dado item para uma variável de previsão de valor real e a descoberta de relações funcionais entre variáveis conforme (FAYYAD et al., 1996).

2.2.3 Associação

As Regras de Associação identificam as relações entre as variáveis ou atributos dos dados, regras que são descritas em uma lista. A descoberta de associação abrange a busca por itens que frequentemente ocorram de forma simultânea em transações do banco de dados (GOLDSCHMIDT; PASSOS; BEZERRA, 2015).

2.2.4 *Clustering* ou Agrupamento

Conforme Fayyad et al. (1996), agrupamento é uma tarefa comum onde se procura identificar um conjunto finito de categorias ou agrupamentos para descrever os dados. Por isso é chamado de aprendizado não supervisionado, ou seja, possui auto-expertise para treinar ou aprender sobre os dados sem rótulos de classe e podendo gerar tais rótulos posteriormente.

2.2.5 Sumarização

A sumarização é a especulação ou reflexo da informação. Muitas informações significativas são desconectadas e resumidas, formando um pequeno conjunto que fornece um diagrama geral de informações (RAJA et al., 2022).

2.3 COVID-19

O surto do vírus SARS-CoV-2 se espalhou rapidamente por todo o mundo após o primeiro caso relatado no final de dezembro de 2019 e foi declarada uma pandemia global pela Organização Mundial da Saúde - OMS (*Organization World Health - WHO*) em 11 de março de 2020 (ALODAT, 2021). A COVID-19 afetou a saúde humana de forma generalizada, alcançando praticamente todos os países. Em 11 de março de 2021, um ano após a declaração do estado de pandemia, haviam mais de 118 milhões de casos positivos notificados no mundo e mais de 2,5 milhões de mortes. Foi necessário um grande esforço global para interromper a propagação do vírus, enquanto as vacinas eram administradas e os tratamentos, desenvolvidos (PINHEIRO et al., 2021).

A primeira reunião do Comitê de Emergência sobre o surto do novo coronavírus na China, convocada pela OMS, de acordo com o Regulamento Sanitário Internacional (RSI),

ocorreu em 23 de janeiro de 2020. Nessa reunião, não houve consenso sobre se o evento constituía uma Emergência de Saúde Pública de Importância Internacional (ESPII). Na segunda reunião, realizada em 30 de janeiro, foi constatado o crescimento no número de casos e de países que reportaram casos confirmados, o que levou à declaração do surto como ESPII, afirma (CRODA; GARCIA, 2020).

O avanço da pandemia foi gerando uma crise de saúde global. Diante desse contexto, foi desencadeada uma corrida pelo desenvolvimento de uma vacina. Cerca de 200 projetos de desenvolvimento foram registrados na OMS (DOMINGUES, 2021). No entanto, uma nova onda de casos de COVID-19 causados pela variante Delta, altamente transmissível, exacerbou a crise de saúde pública mundial e levou à consideração da necessidade potencial e do momento ideal de doses de reforço para as populações já vacinadas (KRAUSE et al., 2021).

A Organização para a Cooperação e Desenvolvimento Econômico - OCDE (*Organisation for Economic Co-operation and Development - OECD)* alertou sobre o impacto negativo que o Coronavírus trouxe ao mundo em relação ao sistema econômico internacional. Em março de 2020, A OCDE divulgou um relatório que estimou que a crise da COVID-19 foi capaz de reduzir pela metade o crescimento da economia mundial até aquele ano, sendo que os efeitos adversos poderiam continuar até 2022 (MAGAZZINO; MELE; MORELLI, 2021).

A utilização de tecnologias de Inteligência Artificial e Aprendizado de Máquina *(Machine Learning - ML)* tem se mostrado cada vez mais importante no combate a pandemias e doenças contagiosas como a COVID-19, onde podem ser armas eficazes contra esse vírus progressivo notoriamente rápido (KANNAN et al., 2020). Algoritmos de aprendizado de máquina desempenham um papel importante na análise e previsão de epidemias (PUNN; SONBHADRA; AGARWAL, 2020). No geral, as técnicas de aprendizagem de máquina vêm sendo inovadoras e úteis, ao prever tendências futuras em casos de COVID-19 e em relação aos aspectos de fatores econômicos.

Nesta dissertação, o conjunto de dados utilizado da COVID-19 foi extraído da *Our World in Data COVID-19* e definidas as seguintes variáveis descritas para análise, de acordo com o repositório do *github owid/covid-19-data*[1], são elas:

- **ISO Code/Código ISO:** ISO 3166-1 alfa-3, código único de três letras para identificação de um país.
- **Continent/Continente:** Continente da localização geográfica.
- **Country/País:** Nome de cada país.

[1] https://github.com/owid/covid-19-data/tree/master/public/data

- **Total Cases/Total de Casos:** Total de casos confirmados de COVID-19. As contagens podem incluir casos prováveis, quando relatados.
- **Total Cases per Million/Total de Casos por Milhão:** Total de casos confirmados de COVID-19 por 1.000.000 de pessoas. As contagens podem incluir casos prováveis, quando relatados.
- **Total Deaths/Total de Mortes:** Total de mortes atribuídas à COVID-19. As contagens podem incluir mortes prováveis, quando relatadas.
- **Total Deaths per Million/Total de Mortes por Milhão:** Total de mortes atribuídas à COVID-19 por 1.000.000 de pessoas. As contagens podem incluir mortes prováveis, quando relatadas.
- **People Vaccinated/Pessoas Vacinadas:** Número total de pessoas que receberam pelo menos uma dose de vacina.

2.4 Indicadores Socioeconômicos e Ambientais

Os Indicadores Econômicos e Sociais são estatísticas que medem diferentes aspectos do desenvolvimento e desempenho da economia e da sociedade (WONG; MICHALOS, 2014). Os indicadores socioeconômicos fornecem uma base para a compreensão do cenário atual de um país, podendo ser fornecidos dados sobre educação, gênero, pobreza, habitação, emprego e outros.

Indicador Ambiental é um valor observado representativo de um fenômeno em estudo, conforme definição dada pela Agência Ambiental Europeia (*European Environmental Agency - EEA*). Estes indicadores quantificam as informações, agregando dados diversos e múltiplos (obtidos de informações confiáveis), podendo ser usados para ilustrar e comunicar fenômenos complexos de forma mais simples, incluindo tendências e progressos ao longo de um determinado período de tempo, de acordo com Herva et al. (2011).

Nesse sentido, os indicadores socioeconômicos e ambientais utilizados neste trabalho foram extraídos do conjunto de dados disponibilizados pelo site *The World Development Indicators*, definidos e descritos cada um conforme informação disponível no Catálogo de Dados[2] (WORLD, 2015). São eles:

- **Population/População:** A População total é baseada na definição de fato de população, que conta todos os residentes, independentemente do status legal ou da cidadania. Os valores apresentados são estimativas semestrais.

[2] https://datacatalog.worldbank.org/search/dataset/0037712/World-Development-Indicators

- **Population Density/Densidade Populacional:** o número de pessoas dividido por área de terra, medida em quilômetros quadrados, do ano mais recente disponível.
- **Median Age/Idade Média:** Idade média da população; projeção da ONU para o ano de 2020.
- **Aged 65 older/Idoso com mais de 65 Anos:** Parcela da população com 65 anos ou mais. foram utilizados os dados mais recente disponíveis para cada país.
- **Aged 70 older/Idoso com mais de 70 Anos:** Parcela da população com 70 anos ou mais; dados relativos ao ano de 2015.
- **Extreme Poverty/Extrema Pobreza:** Parcela da população que vive em extrema pobreza; ano mais recente disponível desde 2010.
- **Cardiovasc Death Rate/Taxa de Mortalidade Cardiovascular:** Taxa de mortalidade por doença cardiovascular em 2017 (número anual de mortes por 100.000 pessoas).
- **Diabetes Prevalence/Prevalência de Diabetes:** Prevalência de diabetes (% de pessoas com idade entre 20 a 79 anos). Com diabetes tipo I ou tipo II. É calculado ajustando-se a uma estrutura etária padrão da população.
- **Handwashing Facilities/Instalações para Lavar as Mãos:** Percentagem da população com instalações básicas para lavagem das mãos nas suas casas. Foram utilizados os dados mais recente disponíveis para cada país.
- **Hospital Beds per Thousand/Leitos Hospitalares por Mil:** Leitos hospitalares por 1.000 pessoas; ano mais recente disponível desde 2010. Leitos Hospitalares incluem leitos de internação disponíveis em hospitais públicos, privados, gerais e especializados e centros de reabilitação.
- **Life Expectancy/Expectativa de Vida:** Expectativa de vida ao nascer, dados de 2019.
- **Human Development Index/Índice de Desenvolvimento Humano:** Um índice composto que mede o desempenho médio em três dimensões básicas do desenvolvimento humano – uma vida longa e saudável, conhecimento e um padrão de vida decente, valores para 2019. Dados disponibilizados pela *Human Development Reports* - UNDP[3].
- **GDP (current US$)/PIB (US$ atual):** O PIB, ou Produto Interno Bruto, é a soma do valor bruto adicionado por todos os produtores residentes na economia,

[3] http://hdr.undp.org/en/indicators/137506

mais quaisquer impostos sobre os produtos e menos quaisquer subsídios não incluídos no valor dos produtos. Os dados estão em dólares americanos correntes.

- **GDP per capita (current US$)/PIB per capita (US$ atuais):** O PIB per capita é o produto interno bruto dividido pela população no mesmo ano. Os dados estão em dólares americanos correntes.
- **Access to Clean Fuels and Technologies for Cooking (% of population)/Acesso a Combustíveis Limpos e Tecnologias para Cozinhar (% da população):** Indica a proporção da população total que usa principalmente combustíveis e tecnologias limpas para cozinhar. De acordo com as diretrizes da OMS, o querosene é excluído dos combustíveis limpos para cozinhar.
- **Access to Electricity (% of Population)/Acesso à eletricidade (% da população):** O acesso à eletricidade é a porcentagem da população com acesso à eletricidade. Os dados de eletrificação são coletados da indústria, pesquisas nacionais e fontes internacionais.
- **Gini Index/Índice Gini:** O índice de Gini mede até que ponto a distribuição de renda (ou, em alguns casos, despesas de consumo) entre indivíduos ou famílias dentro de uma economia, se desvia de uma distribuição perfeitamente igual. Uma curva de Lorenz traça as porcentagens acumuladas da renda total recebida em relação ao número acumulado de beneficiários, começando com o indivíduo ou família mais pobre. Um Índice de Gini de 0 representa igualdade perfeita, enquanto um índice de 1 implica desigualdade perfeita.
- **GNI (current US$)/RNB (US$ atual):** RNB (anteriormente PNB) é a soma do valor adicionado por todos os produtores residentes mais quaisquer impostos sobre produtos (menos subsídios), não incluídos na avaliação da produção, mais as receitas líquidas de renda primária (compensação de empregados e renda de propriedade) do exterior. Os dados estão em dólares americanos correntes.
- **Population in Largest City/População na maior cidade:** é a população urbana que vive na maior área metropolitana do país.
- **Population in Urban Agglomerations of more than 1 million/População em Aglomerações Urbanas de mais de 1 milhão:** é a população do país que vive em áreas metropolitanas que em 2018 tinham uma população de mais de um milhão de pessoas.
- **Oil Rents (% of GDP)/Rendas de Petróleo (% do PIB):** Indica a diferença entre o valor da produção de petróleo bruto a preços regionais e os custos totais de produção.

- **Urban Population/População urbana:** Refere-se às pessoas que vivem em áreas urbanas conforme definido pelos Institutos Nacionais de Estatística. É calculado usando estimativas populacionais do Banco Mundial e proporções urbanas das Perspectivas de Urbanização Mundial das Nações Unidas. A agregação da população urbana e rural pode não totalizar a população total devido às diferentes coberturas dos países.
- **CO_2 emissions (kt)/Emissões de CO_2 (kt):** As emissões de dióxido de carbono são aquelas provenientes da queima de combustíveis fósseis e da fabricação de cimento. Eles incluem dióxido de carbono produzido durante o consumo de combustíveis sólidos, líquidos e gasosos e queima de gás.

3 Revisão Sistemática da Literatura - RSL

A Revisão Sistemática da Literatura (RSL) é um termo genérico, que compreende todos os trabalhos publicados que oferecem um exame da literatura abrangendo assuntos específicos (GALVAO; RICARTE, 2019). Revisar a literatura é essencial para o desenvolvimento de trabalhos acadêmicos e científicos.

Esta RSL foi realizada através da diretrizes da ferramenta online Parsifal[1], abrangendo as seguintes definições: Objetivos e Questões de Pesquisa, Seleção de Fontes de Pesquisa, Critérios de inclusão e exclusão de artigos, Processo de Seleção e Estratégia de Extração de Dados.

3.1 Objetivos e Questões de Pesquisa

A RSL tem como objetivo realizar uma pesquisa referente ao período de 2018 a 2022, para encontrar artigos que tenham aplicado alguma técnica de aprendizado de máquina no contexto da pandemia da COVID-19, correlacionando com indicadores socioeconômicos. E tem como perguntas norteadoras da pesquisa as seguintes questões:

- QP1 - Quais os perfis socieconômicos que influenciam os elevados números de mortes por COVID-19?
- QP2 - Quais os indicadores socioeconômicos mais relevantes em relação a saúde pública?
- QP3 - Quais as técnicas de aprendizado de máquinas mais utilizadas em análises socioeconômicas?
- QP4 - Quais os softwares de análise de dados utilizados para aplicação de técnicas de aprendizagem de máquina no contexto da pandemia da COVID-19?

3.2 *String* de Busca

A pesquisa realizada obteve a coleta de dados utilizando uma *string* de busca avançada, aplicada em várias bibliotecas de indexação de artigos científicos, por meio da seção de condução da ferramenta Parsifal, através de palavras-chaves consultadas em inglês, exportando os arquivos no formato BibTex[2]. A *string* de busca utilizada encontra-se

[1] https://parsif.al/ - é uma ferramenta online desenvolvida para apoiar pesquisadores na realização de revisões sistemáticas de literatura no contexto da Engenharia de Software.

[2] http://www.bibtex.org/ - é uma ferramenta e um formato de arquivo que são usados para descrever e processar listas de referências, principalmente em conjunto com documentos LaTeX.

na Tabela 1.

Tabela 1 – String de Busca

Idiomas	**String de Busca**
Inglês	(”Coronavirus”OR ”COVID-19”OR ”SARSCOV-2”OR ”new coronavirus”OR ”Pandemic” OR ”Public Health”) AND (”Analysis”OR ”socio-economic Indicators”OR ”GDP”OR ”Gini Index”OR ”Human development Index” OR ”PIB”OR ”Per capita income”) AND (”Machine Learning”OR ”Correlation”OR ”Deep learning”OR ”Software”)

Após a exportação dos resultados, por meio do BibTex, os mesmos foram importados para a plataforma Parsifal, e cada artigo foi classificado de acordo com sua relevância em relação às questões de pesquisa.

3.3 Seleção de Fontes de Pesquisa

A definição da busca sistemática foi com base na seleção de fontes de pesquisa. Foram escolhidas bases de artigos que permitam o acesso livre aos artigos completos, sejam artigos *open-access* ou fontes que sejam acessíveis através de convênios da instituição.

Foram escolhidas 5 bases de dados científicos em âmbito internacional nesta pesquisa: ACM Digital Library[3], Google Scholar[4], IEEE Digital Library[5], Science@Direct[6] e Scopus[7].

3.4 Critérios de inclusão e exclusão de artigos

Os artigos que satisfizeram os critérios de busca, foram pré-selecionados de acordo com os seguintes critérios de inclusão (CI) e exclusão (CE):

- CI01 - Estudos sobre indicadores socioeconômicos;
- CI02 - Estudos que abordam a COVID-19;
- CI03 - Publicação em Inglês ou Português;
- CI04 - Publicações entre 2018 e 2022.
- CE01 - Estudos anteriores a 2018;

[3] https://dl.acm.org/
[4] https://scholar.google.com/
[5] https://ieeexplore.ieee.org/Xplore/home.jsp
[6] https://www.sciencedirect.com/
[7] https://www.scopus.com/home.uri

- CE02 - Não é artigo científico;
- CE03 - Texto duplicado;
- CE04 - Texto pouco relevante ao tema.

3.5 Processo de Seleção

O processo de seleção da RSL seguiu através da leitura detalhada dos artigos, com objetivo de garantir que os trabalhos selecionados sejam bastante relevantes ao tema abordado. Com isso foi realizada uma análise através do título e resumo dos artigos pré-selecionados, consequentemente excluindo os menos relevantes para o objetivo deste trabalho. Em seguida, foram filtrados os artigos através da introdução e conclusão, para melhor entendimento, quando não se obteve clareza no título e resumo.

De acordo com os processos da busca sistemática, foram selecionados, a princípio, 814 artigos relacionados ao tema desta RSL (a. Resultado da busca sistemática)[8]. Na primeira filtragem, realizada através da leitura do título e do resumo destes trabalhos, foram escolhidos 91 artigos que mais se adequavam ao tema (b. Resultado da aplicação dos critérios de exclusão)[9], sendo que após essa análise somente 24 artigos se enquadraram em todos os critérios de inclusão (c. Resultado da aplicação dos critérios de inclusão)[10]. Por fim, apenas os 10 artigos que se mostraram totalmente ao encontro das questões de pesquisa, propostas inicialmente, foram detalhados e discutidos. (d. Filtragem após o processo de seleção)[11].

3.6 Estratégia de Extração de Dados

Ao finalizar a seleção dos artigos, seguiu-se a etapa de extração de dados, que tem como estratégia de escolha os seguintes itens:

- Título, Autores e Ano da publicação;
- Indicadores socioeconômicos e dados da pandemia de COVID-19 analisados;
- Técnicas de aprendizado de máquina utilizadas;
- Software ou aplicações de análise de dados;
- Resultados e conclusões.

[8] a. https://bit.ly/3KvGu8E
[9] b. https://bit.ly/3voWYve
[10] c. https://bit.ly/3knMrtM
[11] d. https://bit.ly/3y3K07M

3.7 Critérios de Qualidade

Os artigos foram avaliados de acordo com os critérios de qualidade por meio de 5 questões de avaliação, onde às respostas são atribuídas as três possíveis notas: Sim (1,0), Parcialmente (0,5) e Não (0,0). Questões essas, apresentadas em seguida:

- QA1 - O estudo apresenta objetivos de pesquisa bem definidos?
- QA2 - O estudo identifica os indicadores socioeconômicos de forma clara?
- QA3 - O estudo informa os números de casos/mortes por COVID-19?
- QA4 - O estudo apresenta métodos, técnicas ou ferramentas de aprendizagem de máquina para auxiliar na abordagem?
- QA5 - O estudo avaliado apresenta uma comparação dos estudos com outros trabalhos relacionados ou algum tipo de experimento?

As perguntas de qualidade e as notas de avaliação, foram atribuídas a cada artigo, por fim, as notas foram somadas para possibilitar uma comparação entre os estudos. Pode-se então perceber o grau de coerência dos artigos selecionados para resolução das questões norteadoras desta pesquisa.

3.8 Resultados e Discussões

Depois de analisados os artigos e selecionados os 10 artigos referente à Tabela 2, mais relevantes ao tema abordado, percebeu-se que, mesmo sendo filtradas as publicações no período de 2018 à 2022, restaram apenas trabalhos no período de 2020 à 2021, conforme Tabela 3. Como pode-se perceber poucos trabalhos na literatura estudaram sobre uma possível relação entre a COVID-19 e indicadores socioeconômicos e ambientais, sendo encontrado somente um estudo (A1) que alcançou muita similaridade ao tema abordado.

Como respostas para as 4 questões de pesquisas definidas inicialmente, destaca-se:

QP1 - Quais os perfis socieconômicos que influenciam os elevados números de mortes por COVID-19?

Foi identificado que destaca-se como característica principal do perfil socioeconômico o Produto Interno Bruto - PIB. Assim como os indicadores climáticos, como a poluição do ar ($PM_{2,5}$ e NO_2), assim como as temperaturas mínima e média estão extremamente correlacionados com as mortes por COVID-19 (artigos A3, A4, A5, A7, A9 e A10). O artigo A3 investigou variáveis climáticas e a poluição do ar em países europeus, associando a poluição do ar por NO_2 e a mortalidade por COVID-19. O A4 ressaltou em sua análise a relação entre os poluentes e os casos de morte por COVID-19. O A5 e o A7

Tabela 2 – Notas de Avaliação dos Artigos com Critérios de Qualidade

Título	**QA1**	**QA2**	**QA3**	**QA4**	**QA5**	**Total**
Predicting COVID-19 Spread Level using Socio-Economic Indicators and Machine Learning Technique	1,0	1,0	1,0	1,0	1,0	5,0
Prediction of COVID-19 cases using machine learning for effective public health management	1,0	1,0	1,0	0,5	1,0	4,5
Satellite date and Machine learning reveal a significant correlation between NO_2 and COVID-19 mortality	0,5	0,5	1,0	0,5	1,0	3,5
Correlation between temperature and COVID-19 (suspected, confirmed and death) cases based on machine learning analysis	0,5	0,0	1,0	0,5	1,0	3,0
The nexus between COVID-19 deaths, air pollution and economic growth in New York state: Evidence from Deep Machine Learning	0,5	1,0	0,5	0,5	0,5	3,0
The quest for multidimensional financial immunity to the COVID-19 pandemic: Evidence from international stock markets	0,0	1,0	0,5	1,0	0,0	2,5
Pollution, economic growth, and COVID-19 deaths in India: a machine learning evidence	0,0	1,0	0,5	1,0	0,0	2,5
The relationship between renewable energy and economic growth in a time of Covid-19: A Machine Learning experiment on the Brazilian economy	0,0	0,5	0,5	0,5	1,0	2,5
Socio-economic disparities and COVID-19 in the USA	0,5	0,5	1,0	0,0	0,5	2,5
Data analytics to evaluate the impact of infectious disease on economy: Case study of COVID-19 pandemic	0,0	0,5	0,5	0,5	0,5	2,0

destacaram a relação entre as mortes relacionadas por COVID-19, o crescimento econômico e agentes poluentes ($PM_{2,5}$ e NO_2), mostrando que onde houve um crescimento econômico insustentável, também houve um aumento maior desses poluentes. O A9 analisou que as taxas de mortalidade não estão correlacionadas com as métricas socioeconômicas nos casos de morte por COVID-19 nos Estados Unidos. E o estudo A10, traz como resultados a previsão de que o aumento do número de mortes por COVID-19 obteve relevante impacto econômico nas indústrias, refletindo nos preços de ações e impactando no PIB e na taxa de desemprego.

Tabela 3 – Sumarização dos artigos selecionados.

ID	Autores	Técnicas Utilizadas	Software	Contexto
A1	Mihoub et al. (2020)	SVM, MLP, RF	Scikit-learn, Matplotib	Estudar a correlação entre a propagação do COVID-19 e os indicadores socioeconômicos em diferentes países usando técnicas de inteligência artificial
A2	Ahmad et al. (2021)	SSLPNN, GPR	Não especificado	Prever o número de casos de COVID-19 com base em fatores ambientais e não ambientais, usando Aprendizado de máquina em 5 regiões da Ásia
A3	Amoroso et al. (2021)	SVM, MLP, RF	Não especificado	Avaliar a existência de associação estatística entre a exposição a poluentes e a mortalidade por COVID-19
A4	Siddiqui et al. (2020)	Cluster K-means	Weka	Analisar a correlação entre os efeitos da temperatura no COVID-19 em casos de suspeitas, confirmação e mortes em diferentes regiões da China
A5	Magazzino, Mele e Sarkodie (2021)	Artificial Neural Networks (ANNs), Árvore de Decisão	Oryx2 - Apache Spark	Avaliar a relação entre mortes relacionadas ao COVID-19, crescimento econômico e concentração de poluentes (PM_{10}, $PM_{2,5}$ e NO_2) no estado de Nova York
A6	Zaremba et al. (2021)	Regressão Linear e Análise Fatorial	Não especificado	Desmistificar a imunidade financeira em nível de país e a vulnerabilidade 'a pandemia COVID-19
A7	Mele e Magazzino (2021)	Regressão Linear Preditiva	Oryx no Apache	Explorar a relação entre as emissões de poluição, o crescimento econômico e as mortes por COVID-19 na Índia
A8	Magazzino, Mele e Morelli (2021)	ANNs, MLP	Oryx 2.8.0 (software da Hyderabad, Índia)	Examinar a relação entre o consumo de energia renovável e o crescimento econômico no Brasil, na pandemia de Covid-19
A9	Paul, Englert e Varga (2021)	Regressão não linear e Árvore de Decisão	XGBoost	Examinar a prevalência da COVID-19 e a taxa de mortalidade em relação as condições socioeconômicas locais nos EUA
A10	Hyman et al. (2021)	Regressão Linear Múltipla (MLR)	Panda do Python	Desenvolver um modelo de aprendizagem que pode alavancar a análise de dados para avaliar a gravidade da COVID-19 em diferentes países e seu impacto na economia global

QP2 - Quais os indicadores socioeconômicos mais relevantes em relação à saúde pública?

Os indicadores socioeconômicos mais relevantes à saúde pública foram os indicadores associados as atividades turísticas, Índice de Desenvolvimento Humano - IDH, Índice de Gini, Renda per capita, PIB, entre outros (artigos A1, A2, A6, A9 e A10). Os autores A1 e A2 elencam os indicadores IDH e Índice de Gini como os mais influentes. O A6 apresenta o tamanho da população e os baixos níveis de índice de desenvolvimento humano como fatores significativos para doenças infecciosas emergentes e seus efeitos subsequentes na saúde pública. Em A9 e A10, de forma sucinta, os autores descrevem que o crescimento econômico, junto ao crescimento do PIB, em países mais desenvolvidos contribui para controlar o nível de gastos com saúde pública.

QP3 - Quais as técnicas de aprendizado de máquinas mais utilizadas em análise socioeconômica?

As técnicas de aprendizado de máquina que mais foram utilizadas são: SVM, MLP, RF, RNA e Regressão Linear, de acordo com os artigos A1, A2, A3, A5, A6, A7 e A8. Essas técnicas listadas foram as que melhor obtiveram resultados satisfatórios na análise de dados da COVID-19. Dentre todos os artigos, A1 e A3 destacam que o uso do modelo RF resultou em uma melhor precisão e robustez na classificação de dados em relação ao SVM e o MLP. No entanto os artigos A2, A5, A6, A7 e A8, obtiveram um resultado melhor com uso das técnicas de MLP, RNAs e Regressão Linear em várias áreas geográficas.

QP4 - Quais os softwares de análise de dados utilizados para aplicação de técnicas de aprendizado de máquina no contexto da pandemia de COVID-19?

Com o uso de vários softwares e bibliotecas, as mais utilizadas para aplicações das técnicas foram: *Oryx Apache*[12] (artigos A5, A7 e A8), *Scikit-learn* [13], Weka[14], *XGBoost*[15] e a biblioteca *Pandas* do *Python*[16] (artigos A1, A4, A9 e A10). Pode-se observar que as principais ferramentas utilizadas, são as mesmas que se destacam na aplicação de técnicas de aprendizado de máquina nas mais diversas áreas.

[12] http://oryx.io/
[13] https://scikit-learn.org/stable/
[14] https://www.cs.waikato.ac.nz/ml/weka/
[15] https://xgboost.readthedocs.io/en/stable/
[16] https://www.python.org/

4 Trabalhos Relacionados

Em seguida, serão apresentados de forma breve os trabalhos que obtiveram maior relação com as questões de pesquisa elencadas nesta RSL. Conforme as notas atribuídas e descritas na Tabela 2.

Mihoub et al. (2020) propõem estudar a correlação entre a propagação da COVID-19 e os indicadores socioeconômicos de diferentes países usando técnicas de inteligência artificial. Os níveis de propagação foram divididos em três datas diferentes, entre abril e maio de 2020. Para definir o nível de classe de cada país, três classificadores são propostos e testados: *Support Vectors Machines (SVM)*, *Multi-Layer Perceptrons (MLP)* e *Random Forests (RF)*. O resultado de melhor relevância foi obtido pelo classificador RF, com uma medida F igual a 93,85%, para data de 15 de abril de 2020. A ferramenta utilizada para criação do modelo foi o *Python*, com as bibliotecas: *Numpy*, *Pandas*, *Matplotlib* e *Scikit-learn*.

O estudo de Ahmad et al. (2021) teve como objetivo predizer o número de casos de COVID-19 com base em fatores ambientais (temperatura, umidade, velocidade do vento, índice ultravioleta, elevação, índice de qualidade do ar e nível de poluição) e não ambientais (população, densidade populacional, proporção de gênero e índice de desenvolvimento humano). Foi construído um modelo de classificação binária para predição dos casos de COVID-19 usando os algoritmos *Shallow Single-Layer Perceptron Neural Network – SSLPNN* e *Gaussian Process Regression (GPR)*. O modelo com o algoritmo SSLPNN teve um desempenho excelente, prevendo o número de casos de COVID-19 com uma precisão de 99,09% durante treinamento e uma precisão de 99,04% durante o teste.

O objetivo de Amoroso et al. (2021) foi avaliar a existência de associação estatística entre a exposição a poluentes e a mortalidade por COVID-19. Foram coletados três tipos distintos de dados, as concentrações de poluentes recuperadas dos dados S-5p (monitoramento da atmosfera da Terra, da qualidade do ar, do clima e da camada de ozônio), dados climáticos do ERA5 (cobertura completa das variáveis climáticas em toda a Europa) e os dados que caracterizam o contexto socioeconômico, incluindo taxas de mortalidade, coletados em vários repositórios online. Modelos de regressão multivariada como *Random Forest (RF)*, *Multilayer Perceptron (MLP)*, *Support Vector Machine (SVM)* e modelo de Regressão Linear Simples (LR) foram utilizados. Os autores revelaram uma associação estatística significativa entre a poluição do ar por NO_2 e a mortalidade por COVID-19 e um papel significativo desempenhado pelas características sociodemográficas, como o número de enfermeiros, leitos hospitalares e o produto interno bruto (PIB) per capita.

Os autores Siddiqui et al. (2020) analisaram a correlação entre a temperatura e as diferentes situações de casos da COVID-19 (casos suspeitos, confirmados e óbitos). Com o método de aprendizado de máquina baseado em *cluster k-means* e o uso da ferramenta WEKA, foram utilizados conjuntos de dados de diferentes regiões da China (conforme relatório da OMS). O experimento foi realizado com três tendências de análise com *cluster* onde a cidade de *Hubei* (onde surgiu o vírus da COVID-19) obteve o maior resultado em todas, como: Tendência 1: efeito da temperatura nos casos de óbito, obtendo 33,14% mais números de casos de óbitos; Tendência 2: efeito da temperatura em casos confirmados; e Tendência 3: efeito da temperatura em casos suspeitos.

O objetivo do estudo conduzido por Magazzino, Mele e Sarkodie (2021) foi avaliar a relação entre mortes relacionadas à COVID-19, crescimento econômico e concentração de poluentes (PM_{10}, $PM_{2,5}$ e NO_2) no estado de Nova York. Os autores analisam a relação entre os poluentes e o número de mortes por COVID-19 utilizando árvore de decisão. Depois, utilizaram os resultados das equações hiperbólicas da primeira expressão em Redes Neurais Artificiais (RNAs) para gerar um efeito causal a respeito da mudança no crescimento econômico do estado de Nova York.

Diante da pergunta "O que determina a imunidade financeira de um país diante a uma pandemia global?" os autores Zaremba et al. (2021) investigaram o comportamento de 67 mercados de ações em todo o mundo durante a pandemia da COVID-19 em 2020. Com um conjunto de dados multidimensional contendo: fatores de finanças, economia, demografia, desenvolvimento tecnológico, saúde, governança, cultura e lei, aplicaram técnicas de aprendizado de máquina. Obtiveram como resultados que o mercado de ações em países com baixa taxa de desemprego e povoados por empresas com políticas de investimentos conservadores tendem a ser mais imunes à crise de saúde.

Mele e Magazzino (2021) tiveram como objetivo analisar a relação entre crescimento econômico, emissões de poluentes e mortes por COVID-19 na Índia, usando dois métodos: o modelo econométrico, que verificou a relação causal unidirecional entre o crescimento econômico e a emissão de $PM_{2,5}$, CO_2 e NO_2, e a análise de aprendizado de máquina com algoritmo D_2C, que mostrou uma relação direta entre concentração de $PM_{2,5}$ e as mortes de COVID-19.

A relação entre o consumo de energia renovável e o crescimento econômico do Brasil foi examinada no artigo de Magazzino, Mele e Morelli (2021), utilizando *Multilayer Perceptron (MLP)*. Com o experimento foi verificado o uso mais intensivo de energia renovável que poderá gerar uma aceleração positiva do PIB brasileiro, podendo compensar os efeitos prejudiciais da pandemia global da COVID-19.

O principal objetivo do estudo de Paul, Englert e Varga (2021) foi examinar a correlação entre algumas métricas socioeconômicas, densidade populacional, taxa de casos confirmados e taxa de mortalidade em várias regiões dos EUA, com foco nas variáveis

que significativamente estão correlacionadas com a prevalência de COVID-19 e a taxa de mortalidade pela doença. Foi utilizada uma regressão usando conjunto de árvores de decisão para construção de modelo não linear e com o método chamado de valores de *SHAP (SHapley Additive exPlanations)* para calcular a importância de cada variável.

Segundo Hyman et al. (2021) o objetivo principal de seu trabalho foi desenvolver um modelo de aprendizagem que pode alavancar a análise de dados para avaliar a gravidade da COVID-19 em diferentes países no mundo e seu impacto na economia global. Os autores destacam que os casos de mortes de COVID-19 se relacionam à economia, onde o PIB e a taxa de desemprego mostram que o desempenho da economia foi afetado em comparação ao crescimento potencial de um país trimestralmente.

4.1 Diferenças Destacadas

Alguns autores abordaram as etapas de seleção e pré-processamento de dados, etapas essas que são essenciais para análise de técnicas de *Machine Learning*, porém a maioria com objetivo de prever casos confirmados, mortes ou suspeitas da COVID-19, através de diversos modelos supervisionados e não-supervisionados, recorrendo aos principais métodos: classificação, regressão e agrupamento.

A novidade proposta neste trabalho, como diferencial aos demais, é a análise da relação entre os dados da Pandemia da COVID-19 e os Indicadores Socioeconômicos e Ambientais, através do processo do KDD, onde será realizada cada etapa para a descoberta de novos conhecimento da base de dados que será construída. A mineração dos dados será realizada de forma iterativa. Ao contrário de estudos anteriores, esta análise não se concentrará em realizar a predição, por entender que os dados não apresentam classes bem definidas. Também não pretende-se recorrer a um processo de regressão para estimar números de casos e de mortes por COVID-19. O objetivo deste trabalho é encontrar informações que estejam presentes na base de dados, mas que não são facilmente observadas ou deduzidas.

A análise dos dados será composta pelas cinco etapas do KDD: seleção de dados, pré-processamento, transformação, mineração e interpretação/avaliação dos resultados. E como diferencial, na etapa de mineração dos dados, este trabalho apresenta uma tarefa descritiva de correlação, agrupamento e regras de associação, conforme a etapa do KDD para identificação de padrões compreensíveis, válidos, novos e potencialmente úteis a partir da base de dados.

Para isso, foram coletados *datasets* dos mais diversos sites oficiais e independentes dos países selecionados, dados climáticos, dados de concentrações de poluentes, assim como o *dataset* original da *World Health Organization - WHO*, do Banco Mundial, da OCDE, no site oficial *The Word Bank*. Para este trabalho, foi gerado um *dataset* concatenando todas

as informações, utilizando como suporte principalmente duas bases de dados: da COVID-19 disponível em *Our World in Data* e dos Indicadores Socieconômicos e Ambientais do *The World Bank*.

4.2 Percurso Metodológico

Para uma melhor compreensão das sequências de cada etapa percorrida, visando alcançar os objetivos propostos em relação à RSL e a importância da decisão pelo processo do KDD, nesta seção cada etapa será brevemente descrita.

Inicialmente, foram realizadas todas as etapas da RSL, como definição do tema de pesquisa e dos objetivos específicos da análise de dados, formulação das questões de pesquisa, busca e seleção dos estudos relevantes para a RSL, utilizando critérios pré-estabelecidos de inclusão e exclusão, extração de dados dos estudos selecionados e avaliação da qualidade.

Na segunda etapa do percurso, têm-se os trabalhos relacionados que apresentaram maior relação com as questões de pesquisa da RSL. Na terceira etapa do percurso, foram identificados os resultados das análises dos 10 artigos mais relevantes, de acordo com as questões de pesquisa propostas.

Na quarta etapa do percurso, temos o processo do KDD. Inicialmente, a pesquisa estava direcionada para realizar uma análise visando prever a correlação entre dados da COVID-19 e os Indicadores Socioeconômicos e Ambientais. No entanto, ao analisar a quantidade de dados disponíveis, foi sugerida, durante a qualificação, a mudança para uma análise descritiva utilizando o método do KDD. A relação entre a RSL e o KDD é fundamental para o sucesso da pesquisa, permitindo que os resultados da RSL sejam explorados de forma mais aprofundada, identificando possíveis padrões e relações entre as variáveis.

Na última etapa do percurso, foram obtidos os resultados do KDD. Através da etapa de interpretação e avaliação, foi possível identificar padrões nos dados e gerar *insights* que possam ser utilizados para compreender melhor a relação entre as variáveis, por meio da análise de correlação, agrupamento e regra de associação.

Assim, a relação entre a RSL e o KDD no percurso metodológico é importante para garantir que os resultados da pesquisa sejam embasados em uma análise rigorosa e completa dos dados disponíveis, permitindo explorar os dados de maneira mais ampla e complexa, a fim de obter informações mais precisas e confiáveis sobre a pandemia e seus impactos socioeconômicos e ambientais. Como demonstrado na Figura **??** através do Percurso Metodológico.

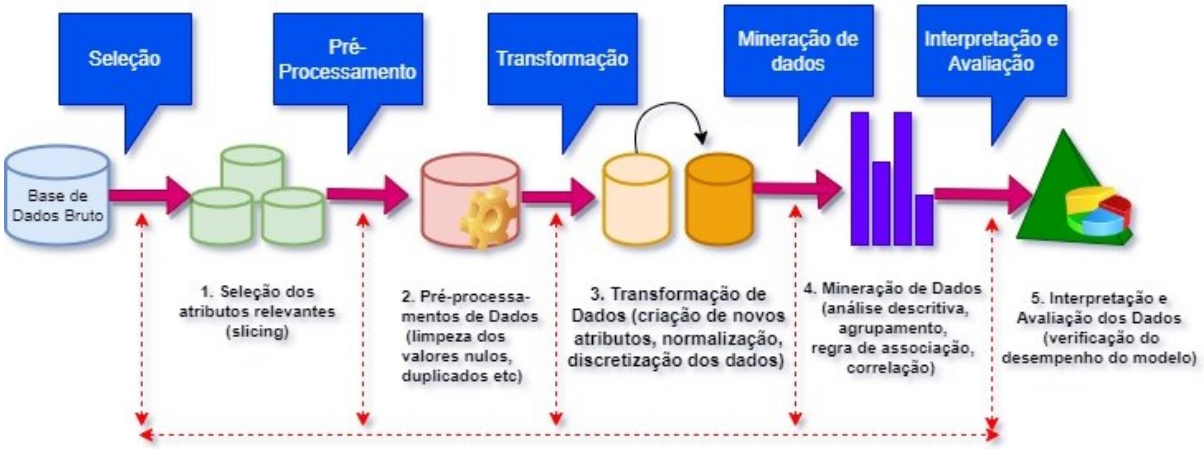
Figura 3 – Etapas do Percurso Metodológico da Pesquisa

Fonte: Própria Autora

5 Aplicação do KDD a base de dados da COVID-19 e Indicadores Socioeconômicos e Ambientais

Neste capítulo são detalhados os métodos que foram aplicados para a análise dos dados da COVID-19 e dos Indicadores Socioeconômicos e Ambientais. A abordagem utilizada pode ser dividida nas cinco grandes etapas do KDD: (1) Seleção de Dados (2) Pré-processamento de Dados (3) Transformação de Dados (4) Mineração de Dados e (5) Interpretação e Avaliação dos Resultados. Para o desenvolvimento do presente estudo, foram seguidas as etapas apresentadas no fluxograma proposto na Figura 4.

Figura 4 – Fluxograma das Etapas do KDD



Fonte: Própria Autora

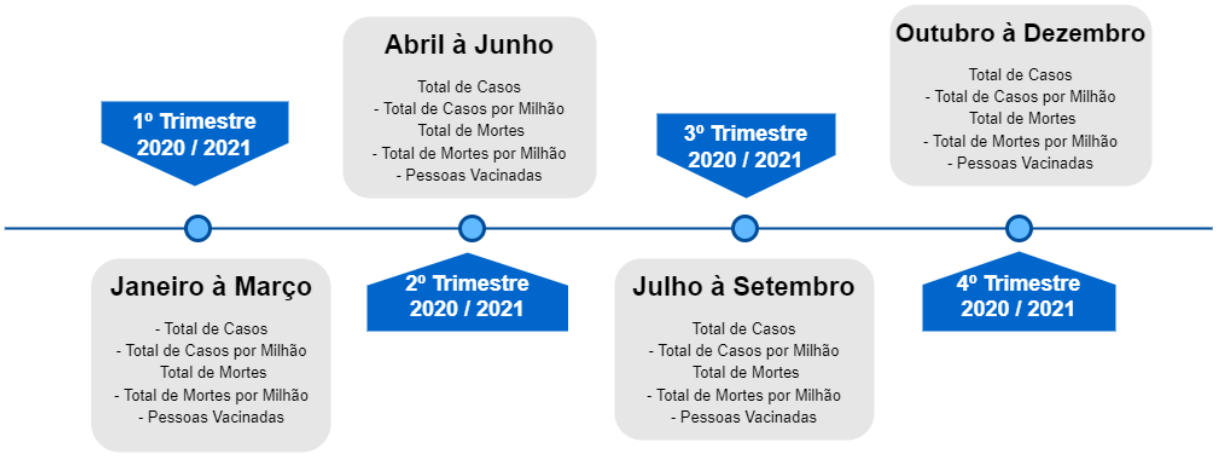
5.1 Base de Dados Brutos

A extração de dados brutos do *dataset* da COVID-19 foram identificadas na fonte oficial disponível no repositório *Our World in Data*[1], foram 39 atributos nos períodos de 2020 a 2021, divididos em 4 trimestres: entre 01 de janeiro a 31 de março, 01 de abril a 30 de junho, 01 de julho a 30 de setembro e 01 de outubro a 31 de dezembro. Os atributos incluem o total de casos, o total de casos por milhão de pessoas, o total de mortes, o total de mortes por milhão de pessoas e pessoas vacinadas como podemos visualizar no recorde temporal na Figura 5, além disso, foram incluídos os atributos da ISO code, continente e nome do país. Para definir o período de coleta, optou-se por selecionar os dados referentes

[1] https://ourworldindata.org/coronavirus

trimestrais de 2020 até 2021, a fim de abranger um período representativo da pandemia e considerar possíveis mudanças ao longo do tempo nos indicadores socioeconômicos e ambientais, vale ressaltar que os dados foram coletados durante a RSL no ano de 2022, sendo assim, os dados de 2021 eram os mais recentes, pois as bases disponibilizam os dados do ano vigente somente no ano posterior.

Figura 5 – Recorte Temporal dos Dados Extraídos Trimestrais da COVID-19



Fonte: Própria Autora

Em seguida, foram coletados os dados dos indicadores socioeconômicos e ambientais na fonte oficial disponível no repositório *The World Bank (DataBank World Development Indicators)*[2], no período de 2020 e 2021, foram extraídos 24 atributos. Portanto, a base completa contém 63 atributos e 198 instâncias, com as descrições dos nomes de cada país. Os dados extraídos de forma estruturada serão armazenados como uma sequência de arquivos de valores separados por vírgula no formato *.CSV*. A coleta de dados para a construção do *dataset* da COVID-19 e dos Indicadores Socioeconômicos e Ambientais foi realizada de forma criteriosa e organizada, garantindo a qualidade e confiabilidade dos dados utilizados na pesquisa. Conforme ilustrado na Tabela 4.

A extração dos dados da COVID-19 e dos Indicadores Socioeconômicos e Ambientais durante os quatro trimestres de 2020 a 2021 é de grande importância para compreender a evolução da pandemia e seus impactos em diferentes contextos sociais e econômicos ao longo do tempo. Além disso, essa abordagem permite identificar possíveis mudanças na dinâmica da pandemia e na relação entre a doença e os indicadores socioeconômicos e ambientais, bem como avaliar a eficácia das políticas públicas e medidas de controle adotadas ao longo do tempo. A análise dos dados em diferentes momentos também permite a identificação de tendências e padrões, bem como a comparação entre diferentes períodos, fornecendo informações valiosas para a tomada de decisão em saúde pública e planejamento de políticas socioeconômicas.

[2] https://databank.worldbank.org/source/world-development-indicators

Tabela 4 – Atributos do Conjunto de Dados Bruto

COVID-19
iso code
continent
country
total case 1st trimester 2020/2021 (31/03)
total case 2nd trimester 2020 /2021 (30/06)
total case 3rd trimester 2020/ 2021 (30/09)
total case 4th trimester 2020/2021 (31/12)
total deaths 1st trimester 2020/ 2021 (31/03)
total deaths 2nd trimester 2020/2021 (30/06)
total deaths 3rd trimester 2020/2021 (30/09)
total deaths 4th trimester 2020/2021 (31/12)
total case per million 1st trimester 2020/2021 (31/03)
total case per million 2nd trimester 2020/2021 (30/06)
total case per million 3rd trimester 2020/2021 (30/09)
total case per million 4th trimester 2020/2021 (31/12)
total deaths per million 1st trimester 2020/2021 (31/03)
total deaths per million 2nd trimester 2020/2021 (30/06)
total deaths per million 3rd trimester 2020/2021 (30/09)
total deaths per million 4th trimester 2020/2021 (31/12)
people vaccinated 1st trimester 2021 (31/03)
people vaccinated 2nd trimester 2021 (30/06)
people vaccinated 3rd trimester 2021 (30/09)
people vaccinated 4th trimester 2021 (31/12)

INDICADORES SOCIOECONÔMICOS E AMBIENTAIS
population
population density
median_age
aged 65 older
aged 70 older
extreme poverty
cardiovasc death rate
diabetes prevalence
handwashing facilities
hospital beds per thousand
life expectancy
human development index
GDP per capita (current us$) in 2019
GDP (current us$) in 2019
access to clean fuels and technologies for cooking (% of population) in 2019
access to electricity (% of population) in 2019
external debt stocks (% of gni) in 2019
GINI index in 2019
GNI (current us$) in 2019
Oil rents (% of GDP) in 2019
population in largest city in 2019
population in urban agglomerations of more than 1 million in 2019
urban population in 2019
co2 emissions (kt) in 2019

5.2 Seleção de Dados

Nesta etapa inicial, após a extração do *dataset* bruto de acordo com a Tabela 4, foi realizada a seleção dos atributos mais relevantes para o processo de mineração de dados, pois geralmente nem todos os dados são necessários. Os dados selecionados do conjunto de dados a serem analisados neste estudo foram obtidos através do procedimento de fatiamento das colunas (*slicing*), selecionando os atributos trimestrais da base da COVID-19 separadamente. Por exemplo, foram selecionados os dados do total de casos por trimestre entre 2020 e 2021, utilizando a função da biblioteca *Pandas*, *iloc*, com o uso da linguagem Python. Esse indexador *iloc* foi útil para realizar a seleção das 8 colunas desejadas por trimestre.

Em seguida, o mesmo procedimento de seleção dos atributos de todos os indicadores, no total de 24 colunas, foi realizado. Após isso, houve a junção dos dois *dataframes* em uma única tabela, por meio do método *pd.concat*, construindo assim a tabela com os dados do total de casos por trimestre selecionados e salvos em um arquivo CSV para análise posterior.

O procedimento foi repetido para as demais colunas trimestrais, criando assim 5 arquivos CSV separadamente com os dados: total de casos por trimestre no período de 2020 a 2021; total de mortes por trimestre no período de 2020 a 2021; total de casos por milhão de pessoas por trimestre no período de 2020 a 2021; total de mortes por milhão de pessoas por trimestre no período de 2020 a 2021 e pessoas vacinadas por trimestre no período de 2021, com a junção dos dados dos indicadores socioeconômicos e ambientais. Consequentemente, foram descartados os atributos *iso code* e *continent*, podendo ser realizada uma análise posteriormente, caso necessário. Já o atributo *country* foi utilizado

como índice através do método *set_index*, onde é substituída a coluna do índice de números pelos nomes de cada país. Os dados estão disponíveis na tabela no formato Excel no *Google Drive*[3], dos atribuídos que foram extraídos das bases.

5.3 Pré-processamentos dos Dados

Antes de qualquer procedimento de mineração, é de extrema importância a etapa de pré-processamento dos dados. Nesta etapa, algumas manipulações de pré-processamento foram realizadas, principalmente tratando os valores nulos/ausentes (*missing*) ou duplicados. Analisando os dados da etapa de seleção que foram gerados anteriormente, a incompletude dos dados foi verificada com a ajuda da função *isnull().sum()*, que calcula a soma dos elementos em cada linha e coluna. Foram identificados vários valores nulos, em média de 13% a 22% de valores *missing* em cada *dataset*. Ao utilizar o método *duplicated()*, não foram encontrados valores duplicados no conjunto de dados.

Em seguida, foram realizados os procedimentos de substituição desses valores nulos com o método *fillna*, preenchendo todos os elementos vazios com a média. Isso foi feito por meio de uma medida estatística, utilizando as medidas de tendência central através da média aritmética por meio da função *np.mean* da biblioteca *numpy*. Após o tratamento dos dados, foi concluído gerando um novo arquivo csv com os dados processados. Todo esse processo foi repetido nos demais arquivos da seleção separadamente, gerando novos arquivos após o pré-processamento.

5.4 Transformação dos Dados

Nessa etapa, os dados foram convertidos em uma forma adaptada para o processo de mineração de dados. Depois que os dados foram processados na etapa anterior, foi realizada a transformação dos dados criando um novo atributo chamado período total para cada arquivo (casos, mortes, casos por milhão, mortes por milhão e pessoas vacinadas), ou seja, foi feita a soma de todos os valores de cada atributo trimestral para o período de 2020 e 2021 utilizando o método *iloc* para selecionar as linhas e colunas desejadas.

Após a análise, verificou-se como os dados estavam distribuídos através da função *describe()*, que apresentou estatísticas descritivas de todas as variáveis, como a quantidade de valores, a média, o desvio padrão, o valor mínimo, os quartis da distribuição e o valor máximo. Foi possível visualizar alguns detalhes de cada dado numérico distintos e encontrar valores discrepantes.

Antes do processo de normalização dos dados, foi realizada uma análise da distribuição dos dados por meio da medida de assimetria (*skewness*). Pôde-se perceber que

[3] https://clck.ru/32qwuX

a distribuição estava enviesada, com alguns dados apresentando assimetria positiva ou negativa. Para essa análise, foi utilizada a função *skew()*. Já a medida de curtose (*kurtosis*), representando a relação com o pico da curva de distribuição de frequência, foi analisada por meio da função *kurt()*. A medida de curtose pode apresentar curtose positiva ou negativa, medindo o grau da cauda na distribuição de frequência e explicando quão alto e nítido é o pico central. Foram identificados vários dados fora da distribuição normal (*gaussiana*), o que motivou a necessidade de aplicar a normalização nos dados.

Foi aplicada a normalização nos dados devido às variáveis estarem em escalas diferentes. Essa transformação foi realizada por meio da biblioteca *scikit-learn*, que realiza a escala das variáveis de entrada para o intervalo 0 e 1 separadamente. Para isso, foi definida uma instância *MinMaxScaler* com hiperparâmetros *feature_range=(0,1)*, que é o padrão para normalização variando entre 0 e 1. Em seguida, foi utilizada a função *fit* para treinar o algoritmo com os dados originais, e então aplicado o normalizador através da função *transform* no conjunto de dados, resultando em um novo conjunto de dados normalizados. A fórmula da normalização Min-Max é simples e consiste em escalonar os dados para um intervalo entre 0 e 1.

$$_{Ynorm}= \frac{X - Xmin}{Xmax - Xmin} \tag{5.1}$$

onde se subtrai os valores X pelo valor mínimo e divide-se o resultado pela diferença entre o valor máximo e o valor mínimo. Essa operação resulta em valores entre 0 e 1.

Além disso, foi realizada a discretização dos dados em alguns atributos que precisavam ser categorizados para serem utilizados em determinados modelos. Para essa transformação, utilizou-se a biblioteca *matplotlib.pyplot*, com a função *pd.cut* do *Pandas*. Foi utilizado o histograma para observar a frequência no eixo y no *range* de valores, onde os intervalos são representados pelos bins. Os bins são os intervalos de discretização dos dados contínuos dentro de uma observação. Todos esses procedimentos foram realizados em cada arquivo csv gerado na etapa de pré-processamento e salvo separadamente. Em seguida, foram apresentados todos os atributos discretizados, juntamente com as quantidades de valores atribuídos para cada variável. Os resultados obtidos estão apresentados na Tabela 5.

Tabela 5 – Classificação dos Dados Discretizados

Atributos	**Atributos Discretizados**	**Qdte de Valores**
IDH	Muito Baixo	16
	Baixo	35
	Médio	50
	Alto	51
	Muito Alto	46
PIB	Alto	194
	Baixo	4
Idade Média	Alta	116
	Baixo	82
Expectativa de Vida	Longevidade Alta	138
	Longevidade Baixa	59
Leitos Hospitalares	Leitos Disponíveis	11
	Leitos Ocupados	187
Instalações para Lavar as Mãos	Instalações Suficientes	152
	Instalações Insuficientes	46
Taxa de Morte Cardiovascular	Taxa Baixa	170
	Taxa Alta	28
Índice de GINI	Alto	179
	Baixo	19
Período Total de Casos	Casos 1º Trimestre de 2020	191
	Casos 2º Trimestre de 2020	4
	Casos 3º Trimestre de 2020	0
	Casos 4º Trimestre de 2020	1
	Casos 1º Trimestre de 2021	0
	Casos 2º Trimestre de 2021	0
	Casos 3º Trimestre de 2021	0
	Casos 4º Trimestre de 2021	1
Período Total de Mortes	Mortes 1º Trimestre de 2020	185
	Mortes 2º Trimestre de 2020	9
	Mortes 3º Trimestre de 2020	1
	Mortes 4º Trimestre de 2020	0
	Mortes 1º Trimestre de 2021	1
	Mortes 2º Trimestre de 2021	1
	Mortes 3º Trimestre de 2021	0
	Mortes 4º Trimestre de 2021	1
Período Total de Casos por Milhão de Pessoas	Casos por milhão de pessoas 1º Trimestre de 2020	94
	Casos por milhão de pessoas 2º Trimestre de 2020	34
	Casos por milhão de pessoas 3º Trimestre de 2020	23
	Casos por milhão de pessoas 4º Trimestre de 2020	25
	Casos por milhão de pessoas 1º Trimestre de 2021	12
	Casos por milhão de pessoas 2º Trimestre de 2021	5
	Casos por milhão de pessoas 3º Trimestre de 2021	3
	Casos por milhão de pessoas 4º Trimestre de 2021	2
Período Total de Mortes por Milhão de Pessoas	Mortes por milhão de pessoas 1º Trimestre de 2020	119
	Mortes por milhão de pessoas 2º Trimestre de 2020	43
	Mortes por milhão de pessoas 3º Trimestre de 2020	23
	Mortes por milhão de pessoas 4º Trimestre de 2020	10
	Mortes por milhão de pessoas 1º Trimestre de 2021	1
	Mortes por milhão de pessoas 2º Trimestre de 2021	0
	Mortes por milhão de pessoas 3º Trimestre de 2021	1
	Mortes por milhão de pessoas 4º Trimestre de 2021	1
Período Pessoas Vacinadas	Pessoas Vacinadas 1º Trimestre de 2021	195
	Pessoas Vacinadas 2º Trimestre de 2021	2
	Pessoas Vacinadas 3º Trimestre de 2021	0
	Pessoas Vacinadas 4º Trimestre de 2021	1

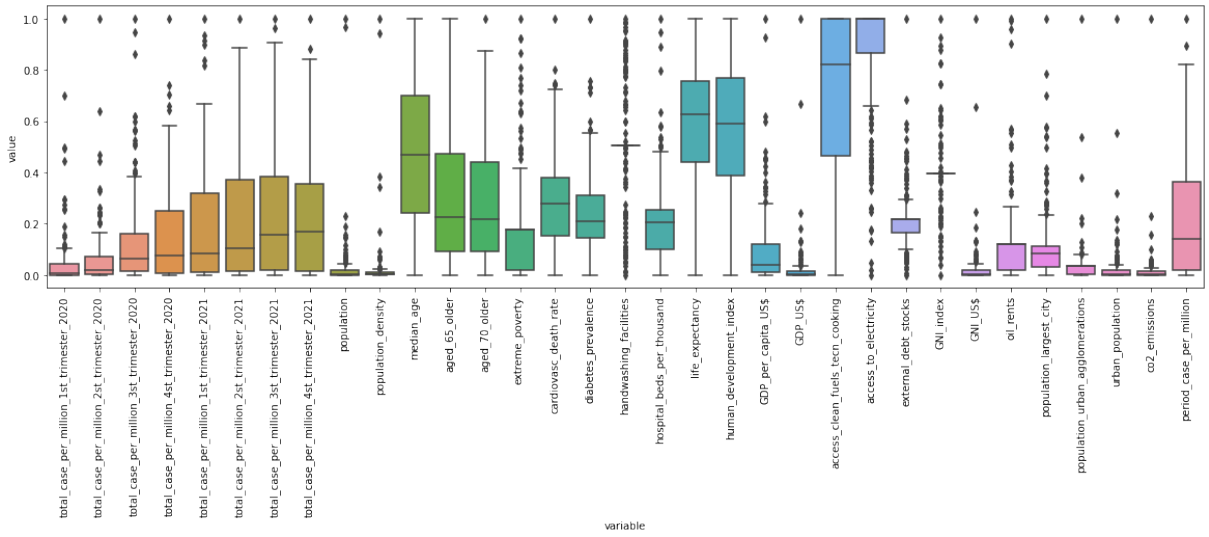
5.5 Mineração dos Dados

Nesta etapa, o principal objetivo é entender a descrição dos dados por meio da sumarização utilizando técnicas como *boxplot*, histograma, correlações e analisar os dados utilizando algoritmos de agrupamento e regra de associação, utilizando as bases de dados

geradas na etapa anterior.

A mineração dos dados iniciou-se com a análise qualitativa por meio do gráfico *boxplot*, que nos permite observar a relação entre as variáveis quantitativas (numéricas), verificando a dispersão, assimetria, média, mediana e *outliers* dos dados. Foi realizada uma análise individual para cada arquivo da etapa anterior de transformação dos dados. Na Figura 6 (por exemplo, no *dataset Total Case per Million*), pode-se observar que algumas variáveis estavam bastante dispersas, assimétricas. Usando a técnica de detecção de *outliers*, com o uso da biblioteca *PyOD import KNN*, foi possível identificar alguns *outliers* em relação ao eixo Y com os valores de cada variável e o eixo X as variáveis. Para plotagem dos gráficos foi necessário renomear algumas variáveis para melhor visualização.

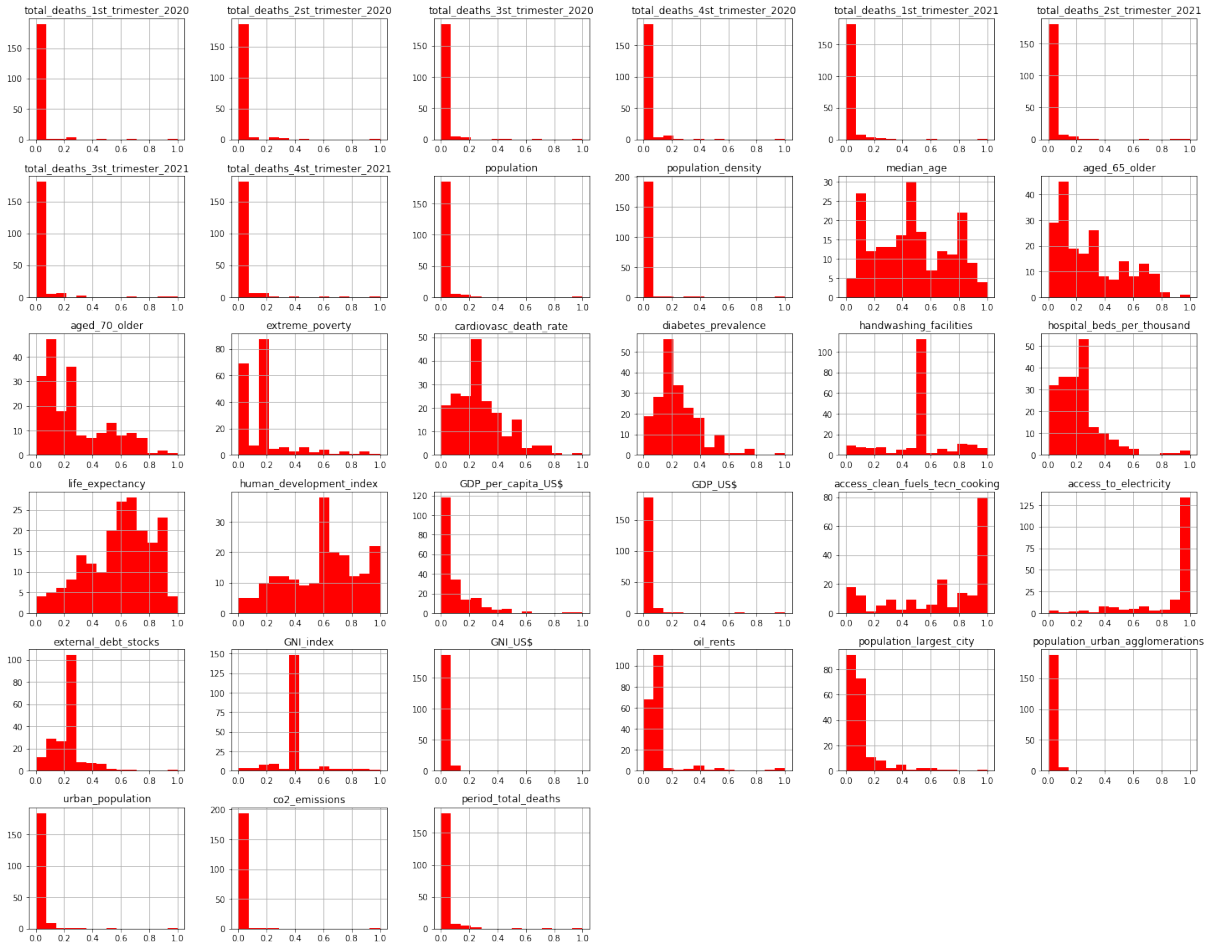
Figura 6 – Boxplot de todas as variáveis do *Dataset Total Case per Million*



Fonte: Própria Autora

Através da análise dos histogramas gerados com o método *hist()* para cada atributo numérico do conjunto de dados, foi possível identificar a presença de assimetria positiva ou negativa na distribuição de algumas variáveis do *dataset total deaths*. Em particular, as variáveis *total deaths* por trimestre, *population, population density, GDP US$, GNI US$, population urban agglomerations, urban population, CO_2 emissions* e *period total deaths* apresentaram uma distribuição enviesada para a direita, com uma concentração acima de 100 mil instâncias no eixo vertical e uma média de probabilidade entre 0,0 e 0,2 no eixo horizontal, que corresponde ao número de casos de mortes. A Figura 7 ilustra esse resultado.

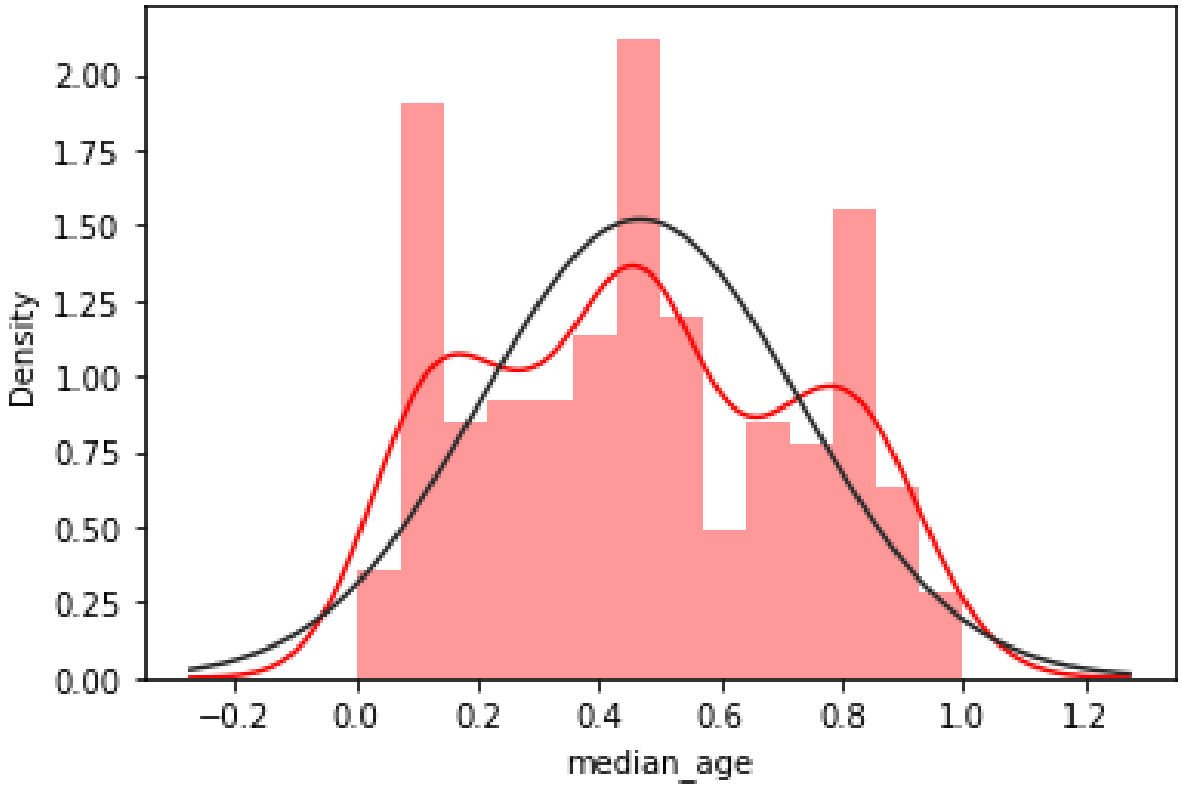
Figura 7 – Histograma *Total Deaths*



Fonte: Própria Autora

Foi realizada outra análise com histograma através do método *distplot()* função da biblioteca *Seaborn* com parâmetro *kde=True* que mostra a forma da distribuição dos dados. Esse método plota a linha de distribuição de probabilidade da normal e gera o número ideal de *bins=K*, calculado através da fórmula de *Sturges* obtendo o número de bins=14. Após o cálculo e a utilização do método *fit=norm* da biblioteca *scipy.stats* e a função *norm*, foi analisada a variável *median age*. Pode-se perceber que o comportamento da distribuição dessa função se assemelha a uma distribuição normal e com uma maior densidade de probabilidade no eixo Y acima de 2.00, no eixo X representa a distribuição dos valores da variável *median age*. Conforme mostrado na Figura 8.

Figura 8 – Histograma da Variável *Median Age*



Fonte: Própria Autora

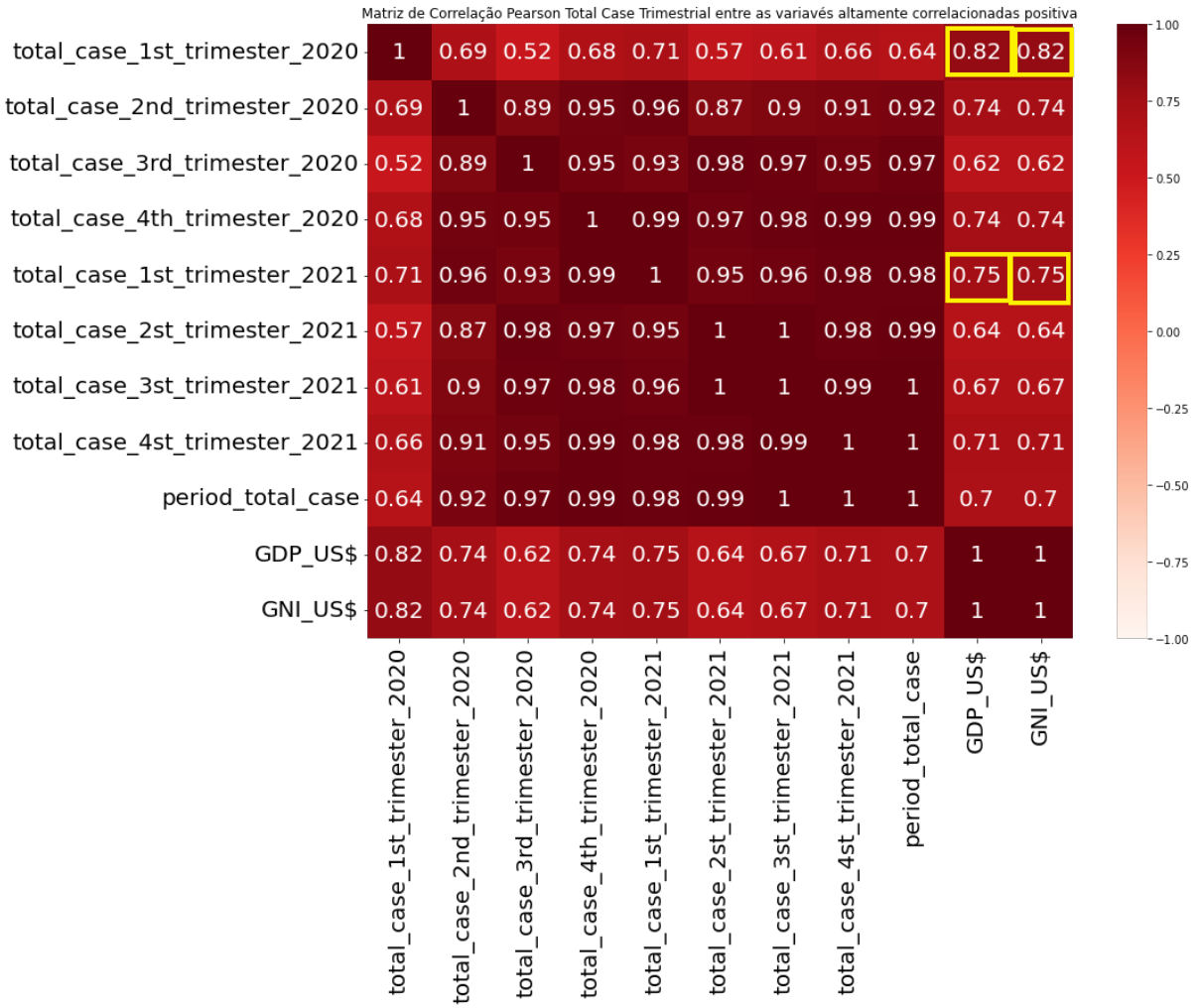
5.5.1 Correlações

Analisando a correlação das variáveis por meio do coeficiente de correlação padrão, chamado de *Pearson (r)*, a interpretação do poder de proporcionalidade compreende valores entre 1 (correlação forte positiva) e -1 (correlação forte negativa, de forma inversa), entre as variáveis. Esse método foi aplicado utilizando a biblioteca *seaborn* e a função *corr()*, considerando variáveis com distribuição próxima da normal e correlação linear. Para visualizar melhor os dados foi utilizada a matriz de correlação, através do gráfico de mapa de calor, gerado com a função *heatmap* da mesma biblioteca. Foram consideradas as *features* relevantes com correlação positiva acima de 60%.

Como exemplo desta análise, foram plotadas somente as variáveis altamente correlacionadas na base do *total case trimester* e as variáveis *total case trimester 2020 e 2021*. Neste caso, é possível identificar forte correlação positiva entre as variáveis *total case trimester 2020 e 2021* e o *period case trimester*, entre as variáveis *GNI US$* e *GDP US$* dos indicadores socieconômicos e ambientais, conforme demonstra o mapa de calor na Figura 9. Não foi encontrada uma forte correlação negativa próximo de -1, entre casos de COVID-19 e os indicadores socioeconômicos e ambientais.

O coeficiente de correlação de *Spearman* é uma medida de correlação não-paramétrica

Figura 9 – Correlação de *Pearson Dataset Total Case Trimester*



Fonte: Própria Autora

que avalia a relação entre duas variáveis contínuas, não-linear e que não seguem uma distribuição específica. Ele é mais robusto em relação a *outliers* e dados não-normais do que o coeficiente de correlação de *Pearson*.

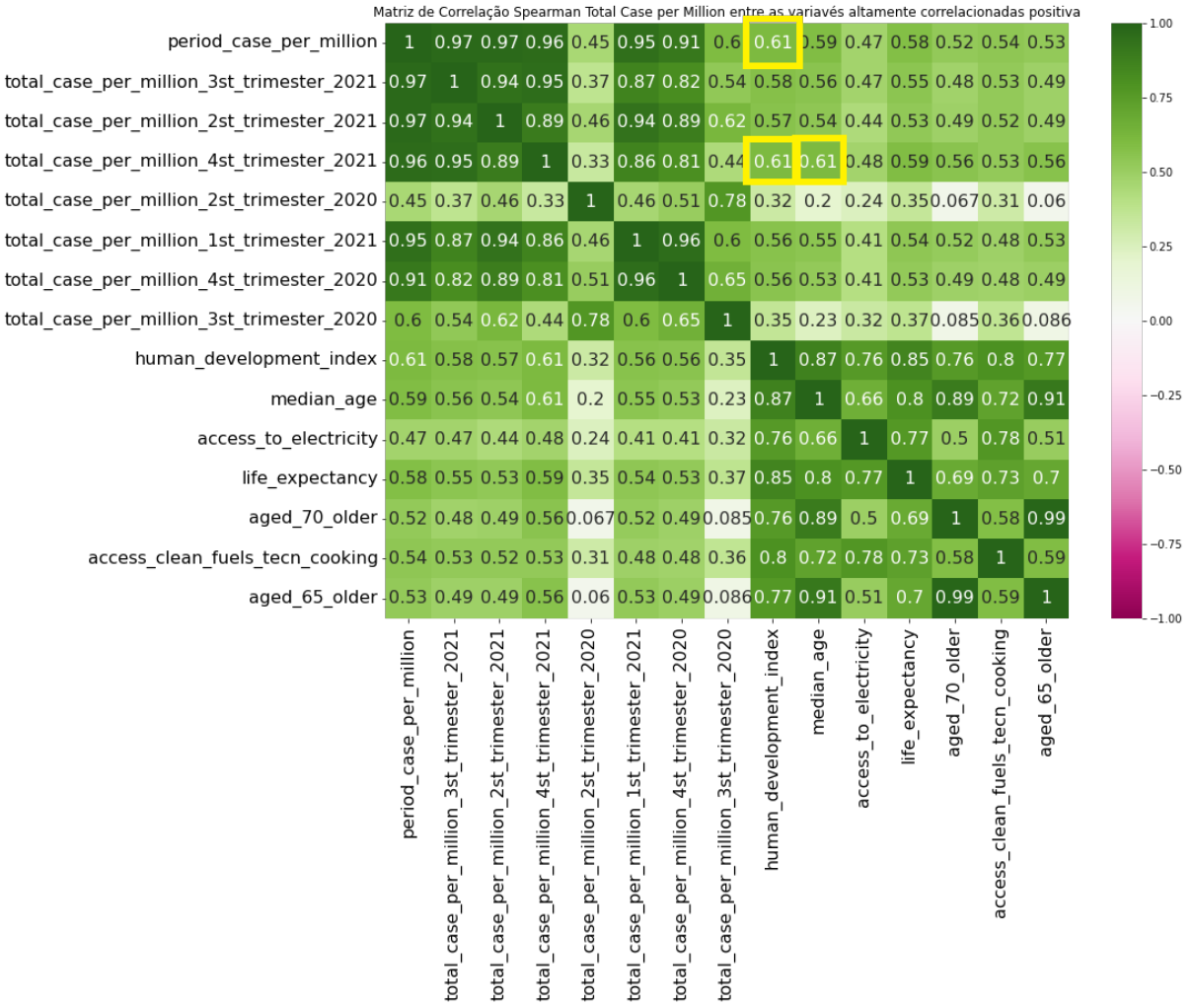
Para calcular o coeficiente de correlação de *Spearman* em dados, basta utilizar o método *corr(method='spearman')* do pacote *seaborn*. Para visualizar a matriz de correlação, pode-se utilizar a função *heatmap* do mesmo pacote, que retorna um gráfico com uma escala de cores indicando o grau de correlação entre as variáveis.

É importante ressaltar que, ao contrário do coeficiente de correlação de *Pearson*, o coeficiente de correlação de *Spearman* não mede apenas correlações lineares entre as variáveis, mas também pode detectar correlações não-lineares. Além disso, ele não assume nenhuma distribuição específica dos dados, tornando-o uma medida mais geral de correlação.

Como exemplo desta análise, foram plotadas somente as variáveis altamente cor-

relacionadas positivamente na base de dados de *total case per million*. Foram definidas as *features* acima de 60%. Podemos identificar correlações dos dados da COVID-19 e os Indicadores Socioeconômicos e Ambientais, que obtiveram forte correlação positiva entre as variáveis: *period case per million, total case per million 4st trimester 2021* de 2020 e 2021, *human development index*, *median age* com 61%. Conforme pode-se perceber na Figura 10. Houve uma correlação muito fraca, ou seja, nenhuma correlação dos casos de COVID-19 e os Indicadores Socioeconômicos e Ambientais entre as variáveis *total case per million 3st trimester 2020* e *aged 70 older*, *aged 65 older* com 0.086%.

Figura 10 – Correlação de *Spearman Total Case Per Million*

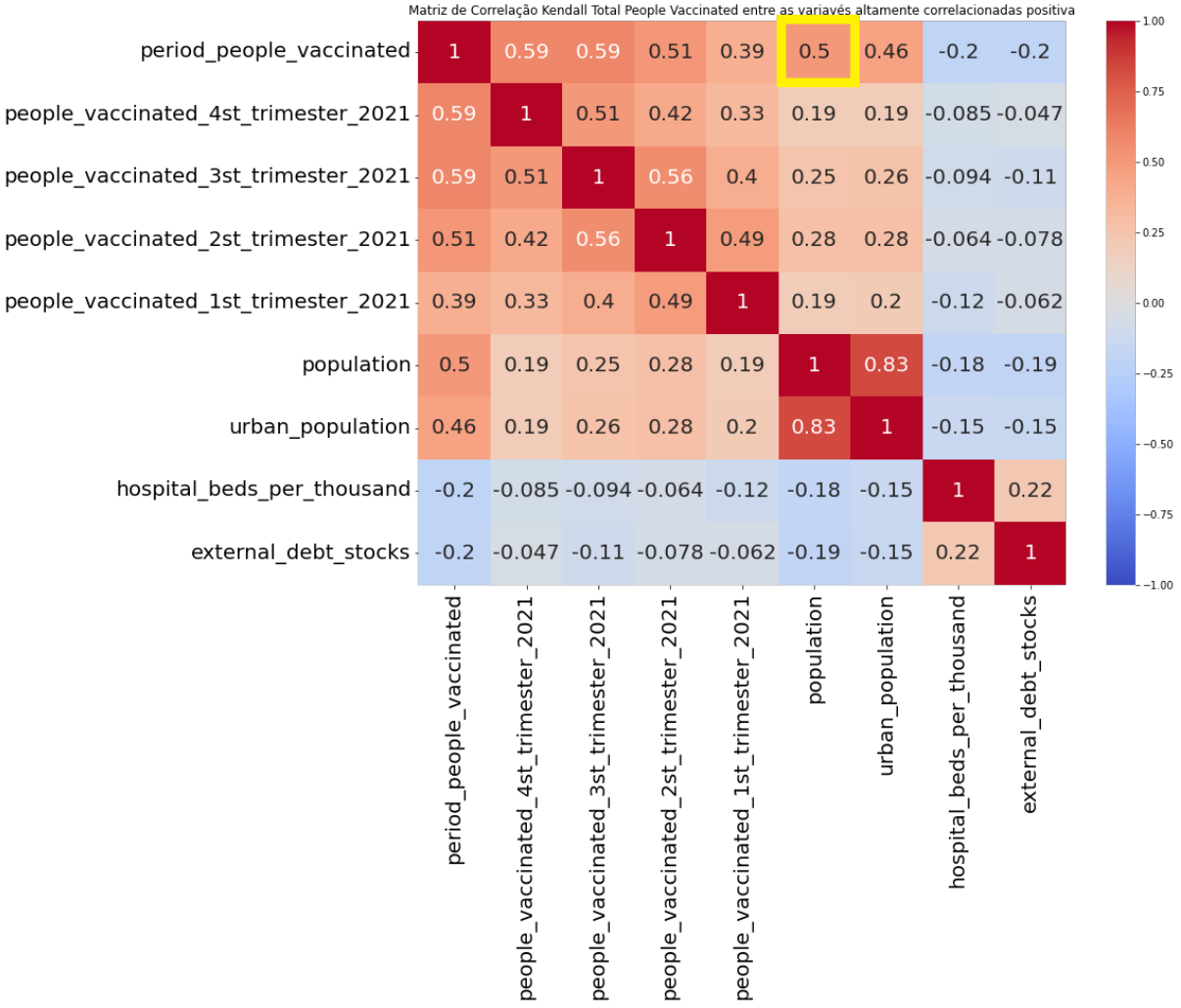


Fonte: Própria Autora

A análise realizada através do coeficiente de correlação de *Kendall*, chamado de *T (tau)*, através do método *corr(method='kendall')*, tem um propósito muito semelhante ao coeficiente de *Spearman*, ambos interpretam relações não-lineares e que não seguem uma distribuição específica, não-paramétrica. No entanto, o nível de significância do *Kendall* torna-se mais confiável em relação ao *Spearman*, o que tende a ter valores menores.

Para visualizar a matriz de correlação, foi utilizado novamente a função *heatmap* do pacote *seaborn*. Em relação ao *period people vaccinated* em 2021, pode-se perceber uma correlação forte positiva com as variáveis *population* e *urban population*, com valores acima de t=0,50. Porém, com valores de correlações negativas, tem-se a variável *external debt stocks*, que se aproxima mais de -1, chegando a t=-0,20, conforme a Figura 11.

Figura 11 – Correlação de *Kendall People Vaccinated*



Fonte: Própria Autora

5.5.2 Agrupamento

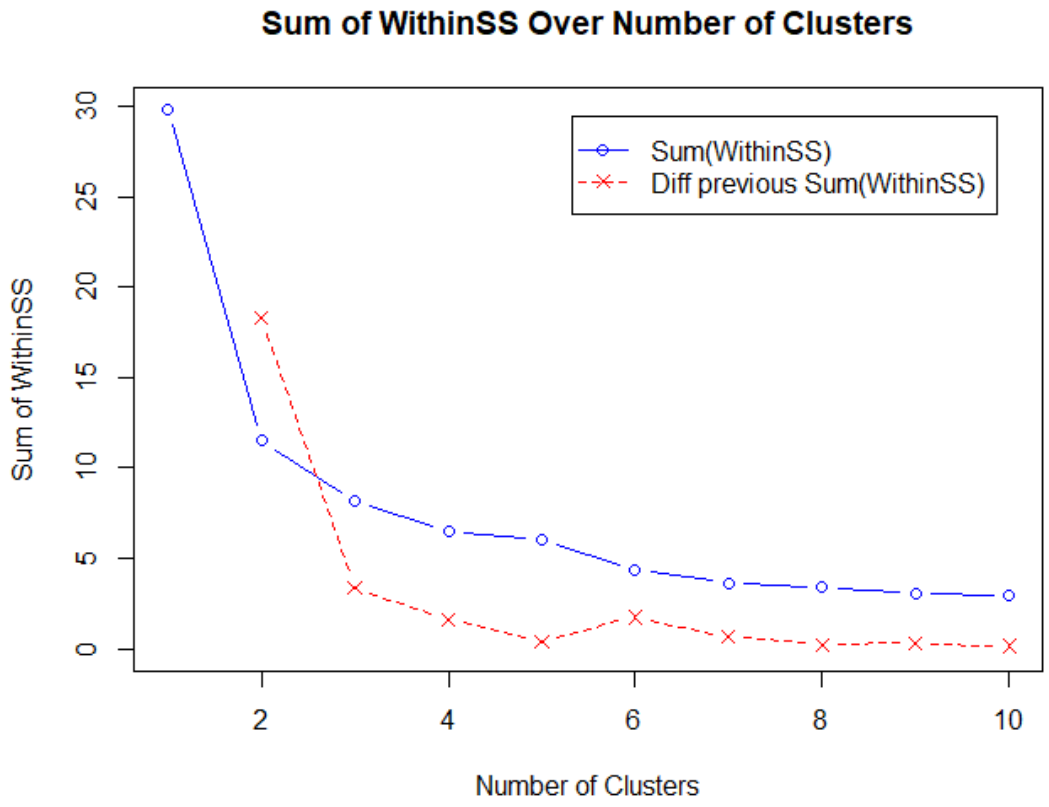
Para essa análise de *Clustering* (Agrupamento), que tem interesse em identificar grupos mais significativos de elementos com determinado grau de semelhança e diferenciando os objetos com uma certa característica em comum entre os grupos, é necessário calcular diversas distâncias entre os elementos, levando em consideração dois tipos de algoritmos: não-hierárquicos e hierárquicos. Nesta análise, foi utilizado o algoritmo *K-means*, que identifica uma coleção de k *clusters* usando uma busca heurística, selecionando centróides

de forma aleatória, reagrupando os dados de forma iterativa de acordo com a distância até os centroides, que são recalculados a cada iteração.

Para esta análise, foram utilizadas algumas variáveis: *population, median age, life expectancy, human development index* e a soma trimestral de casos (*period total case*). Devido à limitação da ferramenta de plotagem do gráfico de dispersão para análise posterior, foi escolhido um total de 5 variáveis para análise.

Inicialmente, foi determinado o número de *clusters* usando o método do cotovelo (*Elbow*), que consiste em calcular a soma dos quadrados *intra-cluster* (ou *Within Clusters Sum of Squares - WCSS*). Para isso, foi utilizado o pacote *Rattle* e o comando *iterate clusters*, que constrói iterativamente mais *clusters* e mede a qualidade de cada modelo. O valor de k variou de 1 a 10 e, a partir do gráfico obtido, pode-se observar que o ponto de inflexão ocorre no valor de k = 3, indicando que o número ótimo de *clusters* para esses dados é 3. Em seguida, foi plotado um gráfico entre o número de *clusters* e o WCSS, como mostra a Figura 12.

Figura 12 – Método Cotovelo no *Dataset Total Case Trimester*



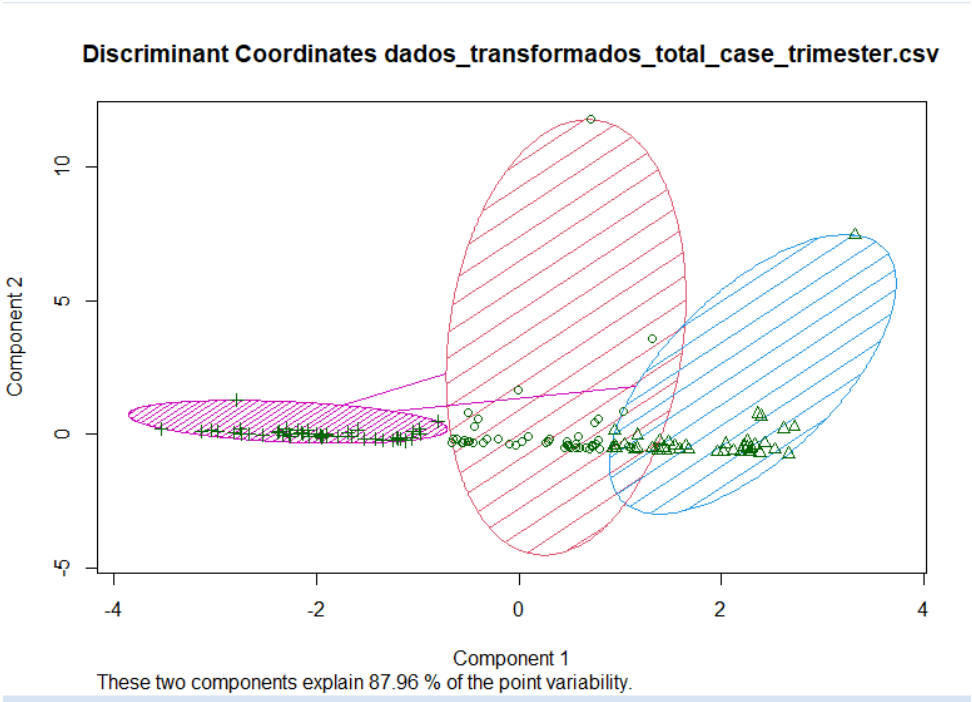
Fonte: Própria Autora

Em seguida, utilizando a função *discriminant plot*, foi gerado um gráfico que mostra como os dados estão distribuídos nos *clusters*, fornecendo as coordenadas discriminantes

que mostram as principais diferenças entre os três grupos distintos, com características semelhantes com algumas sobreposições entre eles. É possível observar também alguns objetos distantes possíveis *outliers* na base *total case trimester*, as características estão representada nos elementos de cada grupo por triângulos, bolas e cruzes.

Os dois componentes principais representam uma redução da dimensionalidade dos dados originais, permitindo uma melhor compreensão das relações entre as variáveis. A análise de componentes principais mostrou que a maior parte da variabilidade nos dados pode ser explicada por esses dois componentes principais, totalizando 87,96% da variabilidade. O primeiro componente principal (eixo X) é composto principalmente pelas variáveis *population, median age, life expectancy, human development index*, e representa a variabilidade socioeconômica dos países. Já o segundo componente principal (eixo Y) é representado principalmente pela variável *period total case* e representa a variabilidade dos Casos por COVID-19 nos países, isso significa que esses dois componentes fornecem informações importantes sobre as relações entre as variáveis originais. Conforme ilustrado na Figura 13.

Figura 13 – Agrupamento da Base de Dados *Total Case Trimester com K-Means*

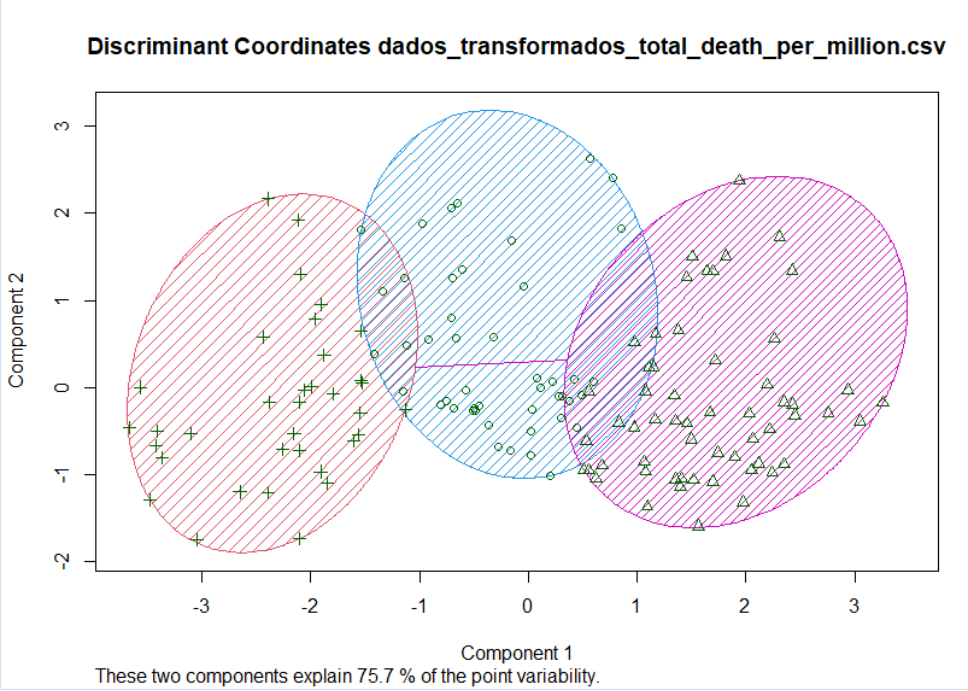


Fonte: Própria Autora

Em uma segunda tentativa de agrupamento, as seguintes variáveis foram utilizadas: *median age, extreme poverty, cardiovascular death rate, life expectancy*, além da variável trimestral de óbitos por milhão (*period death per million*). Foi realizada a mesma análise utilizando o Método do Cotovelo e obteve-se como melhor valor de K=3. Ao utilizar a função *discriminant plot*, foi gerado um gráfico que mostra como os *clusters* estão distribuídos nos

dados. Seguindo o mesmo método, é possível observar as características estão mais dispersas, isso pode indicar que as amostras são mais heterogêneas e com algumas sobreposições. A análise de componentes principais indicaram que dois componentes principais explicam 75,7% da variabilidade total dos dados. O primeiro componente principal (eixo X) é composto principalmente pelas variáveis *median age, extreme poverty, cardiovascular death rate e life expectancy*, e representa a variabilidade socioeconômica dos países. Já o segundo componente principal (eixo Y) é representado principalmente pela variável *period death per million* e representa a variabilidade da mortalidade por COVID-19 nos países. Conforme a Figura 14

Figura 14 – Agrupamento da Base de Dados *Total Death Per Million com K-Means*



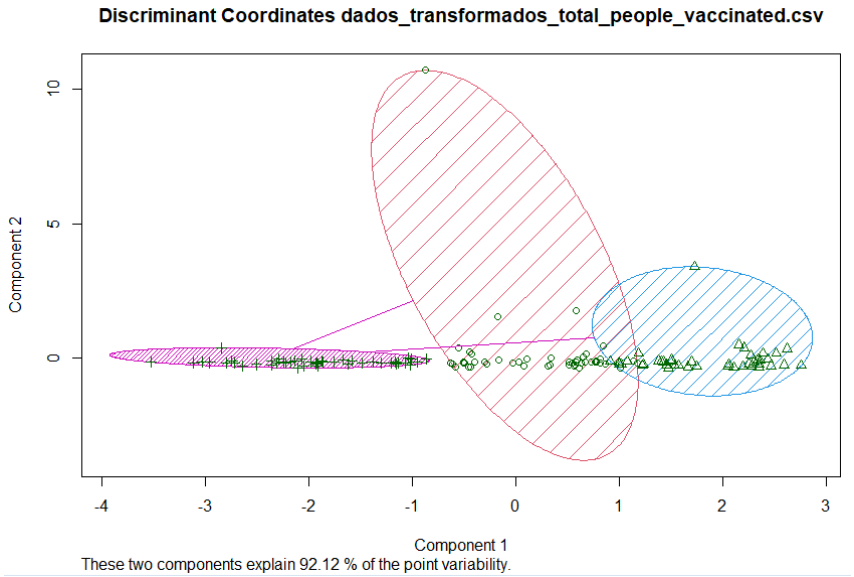
Fonte: Própria Autora

Outra análise foi realizada com as seguintes variáveis: *median age, life expectancy, human development index*, incluindo a variável da soma trimestral (*period total vaccinated*). Novamente obteve-se o resultado de números de *cluster* no Método de Cotovelo com melhor valor de K=3. A análise resultou em um gráfico que exibe como os dados estão distribuídos nos três *clusters*, indicando que os *clusters* estão muito próximos um do outro e há uma pequena sobreposição nas características dos dados, significa que existem áreas em que as amostras são muito semelhantes entre os *clusters* e alguns objetos distantes com possíveis *outliers*.

Ao aplicar o componentes principais nos dados foram identificados dois componentes principais que explicam 92,12% da variabilidade dos dados. O primeiro componente principal (eixo X) é composto principalmente pelas variáveis *median age, life expectancy*

e human development index, e representa a variabilidade socioeconômica dos países. Já o segundo componente principal (eixo Y) é representado principalmente pela variável *period total vaccinated* e representa a variabilidade em que os países que implementaram campanhas de vacinação. Conforme Figura 15. Lembrando que foram realizadas análises com outras variáveis além das utilizadas, porém essas foram que obtiveram melhores resultados.

Figura 15 – Agrupamento da Base de Dados *People Vaccinated com K-Means*



Fonte: Própria Autora

5.5.3 Regras de Associação

Na presente análise, temos uma estrutura de regras de associação, em que um antecedente (*lhs*) e um consequente (*rhs*) são representados por um subconjunto de itens. Cada regra apresenta um conjunto de itens, conhecido como *Itemset*, onde existem relações entre eles. Foram utilizadas as variáveis discretizadas descritas na Tabela 5 para realizar as relações entre elas.

Para identificar o conjunto de itens que apresentam frequência nos dados, foi utilizado o algoritmo *Apriori* por meio do pacote *Rattle* da Linguagem R. Ao selecionar as variáveis de entrada para análise de regras, é obrigatório incluir um identificador (*ident*) exclusivo para cada observação (linha) dos dados, além das variáveis discretizadas. As variáveis categóricas *iso_code, continent* e *country* foram mantidas, e a última foi selecionada como *ident*, como apresentado na tela principal dos dados para análise na Figura 16.

Figura 16 – Visualização das variáveis categóricas para análise de Regras de Associação

Input Ignore Weight Calculator: Target Data Type: Auto Categoric Numeric Survival

No.	Variable	Data Type	Input	Target	Risk	Ident	Ignore	Weight	Comment
1	iso_code	Categoric	○	○	○	○	●	○	Unique: 198
2	continent	Categoric	○	○	○	○	●	○	Unique: 6
3	country	Categoric	○	○	○	●	○	○	Unique: 198
4	IDH	Categoric	○	○	○	○	●	○	Unique: 5
5	PIB	Categoric	○	○	○	○	●	○	Unique: 2
6	idade_media	Categoric	○	○	○	○	●	○	Unique: 3
7	expectativa_vida	Categoric	○	○	○	○	●	○	Unique: 3
8	leitos_hospitalares	Categoric	●	○	○	○	○	○	Unique: 2
9	inst_lavar_maos	Categoric	○	○	○	○	●	○	Unique: 2
10	morte_cardiovascular	Categoric	○	○	○	○	●	○	Unique: 2
11	indice_GINI	Categoric	○	○	○	○	●	○	Unique: 2
12	periodo_total_casos	Categoric	●	○	○	○	○	○	Unique: 5
13	periodo_total_mortes	Categoric	○	○	○	○	●	○	Unique: 6
14	periodo_casos_milhão	Categoric	○	○	○	○	●	○	Unique: 8
15	periodo_mortes_milhão	Categoric	○	○	○	○	●	○	Unique: 7
16	periodo_pessoa_vacinada	Categoric	○	○	○	○	●	○	Unique: 3

Fonte: Própria Autora

Ao utilizar a função *show rules* com as variáveis período total casos e leitos hospitalares, com 2 regras encontradas e 138 transações, foi gerada a regra de associação com critério em grau de *confidence* e *support* com valor padrão mínimo de 0.100 e com número de condições 2 *itemsets*. Todas as análises foram realizadas com essas medidas. Foi encontrada uma associação entre o número de casos da COVID-19 no 1º trimestre de 2020 e o aumento de leitos hospitalares ocupados nesse período, considerando o grau de confiança, suporte e *lift* alto, o que é considerado como a regra mais importante nesse caso, ou seja, as variáveis estão bem mais correlacionadas no conjunto de dados. Isso pode ser observado na Figura 17.

Figura 17 – Regra de Associação das variáveis Período Total Casos e Leitos Hospitalares

```
All Rules

    lhs                                                rhs
[1] {periodo_total_casos=CASOS_1°_TRIMESTRE_2020} => {leitos_hospitalares=LEITOS_OCUPADOS}
[2] {leitos_hospitalares=LEITOS_OCUPADOS}         => {periodo_total_casos=CASOS_1°_TRIMESTRE_2020}
```

Fonte: Própria Autora

Outra análise de associação foi realizada com as variáveis do período total mortes e o índice de GINI, utilizando o mesmo critério de grau de confiança e suporte mínimo e com 2 regras encontradas em um conjunto de 138 transações. Foi observado que no 1º

trimestre de 2020, o índice de Gini foi associado a um alto índice de mortes durante o período, com grau de confiança elevado. Essa análise será avaliada na próxima etapa do KDD. A Figura 18 mostra a regra de associação encontrada.

Figura 18 – Regra de Associação das variáveis Período Total de Mortes e Índice de GINI

```
All Rules

    lhs                                                 rhs
[1] {indice_GINI=ALTO}                               => {periodo_total_mortes=MORTES_1°_TRIMESTRE_DE_2020}
[2] {periodo_total_mortes=MORTES_1°_TRIMESTRE_DE_2020} => {indice_GINI=ALTO}
```

Fonte: Própria Autora

Na análise de associação realizada com as variáveis referentes ao período pessoas vacinadas e o PIB, foram encontradas duas regras com base em 138 transações. Foi possível observar que no 1º trimestre de 2021, o PIB apresentou um alto valor de associação, pois foi o período em que a economia começou a se recuperar após a aplicação das primeiras doses das vacinas. Essa associação apresentou um grau de confiança elevado, como pode ser visto na Figura 19.

Figura 19 – Regra de Associação das variáveis Período Pessoas Vacinas e PIB

```
All Rules

    lhs                                                  rhs
[1] {PIB=ALTO}                                        => {periodo_pessoa_vacinada=PESSOAS_VACINADA_1°_TRIMESTRE_DE_2021}
[2] {periodo_pessoa_vacinada=PESSOAS_VACINADA_1°_TRIMESTRE_DE_2021} => {PIB=ALTO}
```

Fonte: Própria Autora

Realizando a análise com as variáveis do total de mortes por milhão de pessoas e a expectativa de vida, foram geradas 12 regras e 138 transações. É possível perceber que, em países com a expectativa de vida classificada como de longevidade baixa, houve um maior número de mortes por milhão no 1º trimestre de 2020, ou seja, logo no início da pandemia. Enquanto que países com a expectativa de vida classificada como de longevidade alta, as mortes começaram a prevalecer no 2º e 3º trimestre de 2020. Conforme pode-se observar na Figura 20.

Figura 20 – Regra de Associação das variáveis Total de Mortes por Milhão de Pessoas e Expectativa de Vida

```
All Rules

    lhs                                                  rhs
[1] {expectativa_vida=LONGEVIDADE_BAIXA}              => {periodo_mortes_milhão=MORTES_POR_MILHÃO_1°_TRIMESTRE_DE_2020}
[2] {periodo_mortes_milhão=MORTES_POR_MILHÃO_1°_TRIMESTRE_DE_2020} => {expectativa_vida=LONGEVIDADE_BAIXA}
[3] {periodo_mortes_milhão=MORTES_POR_MILHÃO_3°_TRIMESTRE_DE_2020} => {expectativa_vida=LONGEVIDADE_ALTA}
[4] {expectativa_vida=LONGEVIDADE_ALTA}               => {periodo_mortes_milhão=MORTES_POR_MILHÃO_3°_TRIMESTRE_DE_2020}
[5] {periodo_mortes_milhão=MORTES_POR_MILHÃO_2°_TRIMESTRE_DE_2020} => {expectativa_vida=LONGEVIDADE_ALTA}
[6] {expectativa_vida=LONGEVIDADE_ALTA}               => {periodo_mortes_milhão=MORTES_POR_MILHÃO_2°_TRIMESTRE_DE_2020}
[7] {periodo_mortes_milhão=MORTES_POR_MILHÃO_1°_TRIMESTRE_DE_2020} => {expectativa_vida=LONGEVIDADE_ALTA}
[8] {expectativa_vida=LONGEVIDADE_ALTA}               => {periodo_mortes_milhão=MORTES_POR_MILHÃO_1°_TRIMESTRE_DE_2020}
```

Fonte: Própria Autora

Por fim, a análise com as variáveis por período de casos por milhão de pessoas e média de idade gerou 6 regras e 138 transações. Pode-se perceber que no período do 4º trimestre de 2020, houve um aumento dos casos nos países em que a média de idade foi considerada baixa. Enquanto nos demais períodos (1º e 2º trimestres de 2020), houve um número maior de casos onde a média de idade era alta. A regra de associação correspondente está apresentada na Figura 21.

Figura 21 – Regra de Associação das variáveis Período Casos por Milhão de Pessoas e Média de Idade

```
All Rules

    lhs                                                                rhs
[1] {periodo_casos_milhão=CASOS_POR_MILHÃO_4°_TRIMESTRE_DE_2020} => {idade_media=BAIXO}
[2] {idade_media=BAIXO}                                            => {periodo_casos_milhão=CASOS_POR_MILHÃO_4°_TRIMESTRE_DE_2020}
[3] {periodo_casos_milhão=CASOS_POR_MILHÃO_1°_TRIMESTRE_DE_2020} => {idade_media=ALTA}
[4] {idade_media=ALTA}                                             => {periodo_casos_milhão=CASOS_POR_MILHÃO_1°_TRIMESTRE_DE_2020}
[5] {idade_media=ALTA}                                             => {periodo_casos_milhão=CASOS_POR_MILHÃO_2°_TRIMESTRE_DE_2020}
[6] {periodo_casos_milhão=CASOS_POR_MILHÃO_2°_TRIMESTRE_DE_2020} => {idade_media=ALTA}
```

Fonte: Própria Autora

5.6 Interpretação e Avaliação dos Dados

5.6.1 Correlações

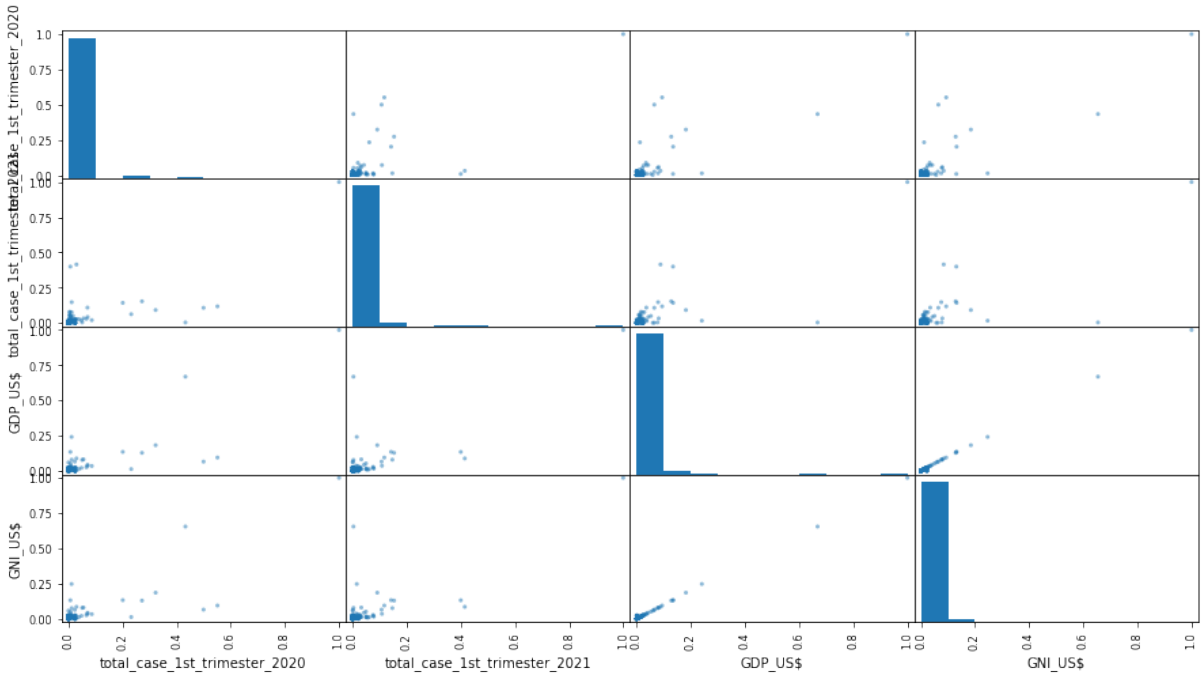
A correlação é uma medida do grau de relacionamento entre duas variáveis que geralmente é um número entre 1 e -1. A magnitude representa a força da correlação e o sinal representa a direção da correlação. Um alto grau de correlação (mais próximo de 1 ou -1) indica que as duas variáveis são altamente correlacionadas, positivamente ou negativamente.

A maneira utilizada para verificar a correlação entre atributos foi a função *scatter matrix* do Pandas conhecida como matriz de dispersão, que plota um gráfico de dispersão para cada par de variáveis e apresenta a correlação entre as duas variáveis que estão sendo comparadas, plotada em um dos eixos, e os pontos representam as observações. Na diagonal principal da matriz é apresentado o histograma que mostra a distribuição de cada variável. A avaliação será realizada em todas as análises da etapa anterior apresentando essa técnica de visualização que ajuda a compreender a natureza das distribuições dos dados.

A correlação de *Pearson* realizada na análise entre as variáveis da base *Total Case Trimester*, apresentada na Figura 9, considerando que existem 11 atributos numéricos, resulta em $11^2 = 121$ plotagens, o que não caberia em uma única página. Portanto, serão utilizados apenas alguns atributos promissores que parecem mais correlacionados. Nessa plotagem, é possível observar uma tendência ascendente e poucos pontos dispersos, onde o *Total Case 1º Trimester* nos períodos de 2020 e 2021 tende a crescer à medida que o

PIB e a RNB - Renda Nacional Bruta (com magnitude de 0,82% e 75%) estão em alta, indicando uma forte correlação. Por motivo da normalização, os dados se concentram mais próximo de zero. Conforme a Figura 22.

Figura 22 – Matriz de Dispersão das Variáveis *Total Case Trimester*

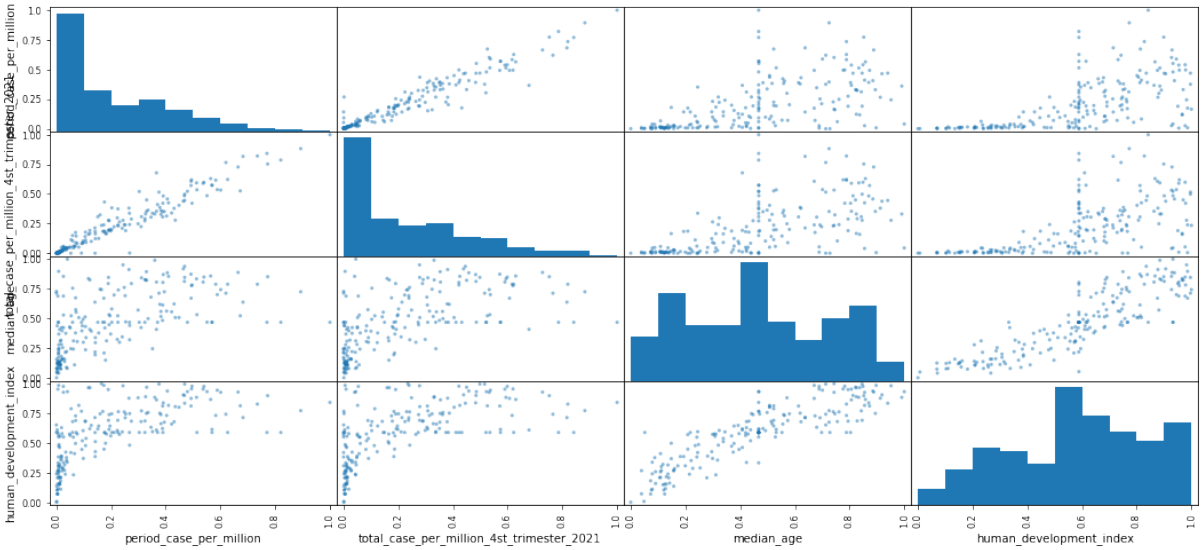


Fonte: Própria Autora

A correlação de *Spearman* entre as variáveis da base *Total Case Per Million* apresentada na Figura 10, uma vez que existem 15 atributos numéricos, é obtida a partir de $15^2 = 225$ plotagens, o que não caberia em uma página. Por isso, foram analisados apenas em alguns atributos promissores que parecem mais correlacionados. Na avaliação, é possível observar uma pequena tendência ascendente e alguns pontos dispersos comuns em correlações não lineares. Indica que, à medida que *aged 65 older* aumenta, os casos no período de *total case per million 3st trimester 2020* tendem a diminuir. O valor dessa correlação foi de 0,085%, indicando uma correlação muito fraca, ou seja, nenhuma correlação. Por isso, não foi apresentado no gráfico de dispersão.

No entanto, o gráfico de dispersão mostra a relação entre as variáveis *total case per million 4st trimester 2021* e *period case per million*, *median age*, e *human development index* com uma correlação de *Spearman* de 0,61%. Pode-se observar que existe uma tendência positiva na relação entre *total case per million 4st trimester 2021* e *median age*, indicando que países com uma população mais velha apresentam uma maior incidência de casos de COVID-19 por milhão de habitantes. Além disso, podemos ver que países com um IDH mais elevado tendem a apresentar uma menor incidência de casos de COVID-19 por milhão de habitantes. Conforme a Figura 23.

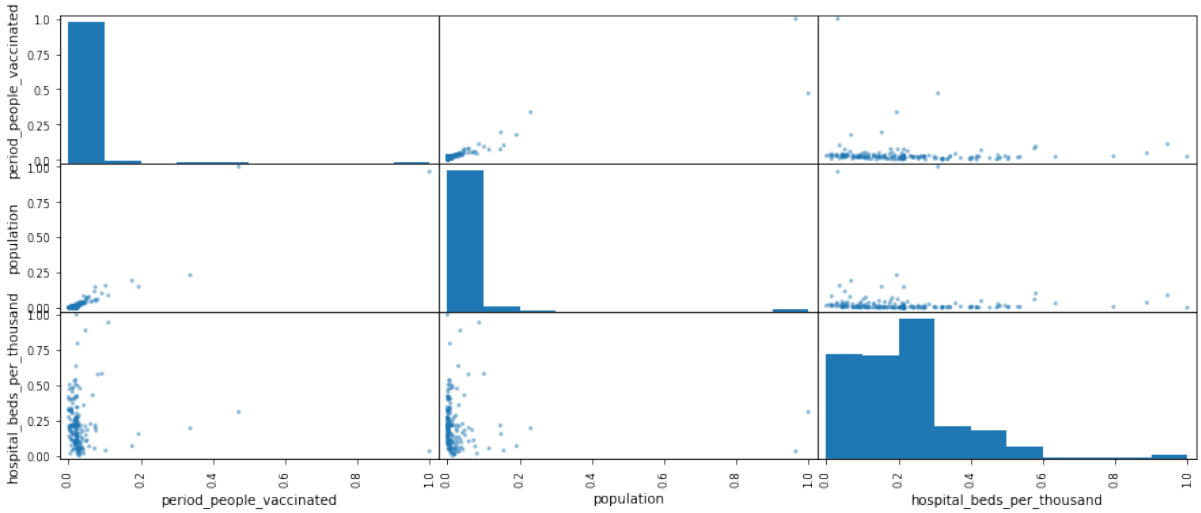
Figura 23 – Matriz de Dispersão das Variáveis *Total Case per Million Trimester*



Fonte: Própria Autora

A correlação de *Kendall* entre as variáveis da base *People Vaccinated*, apresentada na Figura 11, envolve 9 atributos numéricos, o que resultaria em $9^2 = 81$ plotagens. No entanto, apenas alguns atributos que parecem mais correlacionados serão utilizados. Na plotagem entre as variáveis *period people vaccinated* e *population*, é possível observar uma tendência ascendente e alguns pontos dispersos, comum em correlações não lineares. A magnitude da correlação positiva de 0.50 indica uma correlação moderada entre as variáveis, sugerindo que países com maior população tendem a ter mais pessoas vacinadas em termos absolutos. No segundo caso, a correlação de *Kendall* de -0,2 indica uma correlação fraca e negativa entre as variáveis, sugerindo que países com mais leitos hospitalares por mil habitantes podem ter menos casos de COVID-19 por habitantes. Como podemos perceber no gráfico de dispersão da Figura 24.

Figura 24 – Matriz de Dispersão das Variáveis *People Vaccinated*



Fonte: Própria Autora

5.6.2 Agrupamento

Esta avaliação, depois de gerar o modelo de *clustering* apresentado na Figura 13, começa com o tamanho de cada *cluster*, que é simplesmente uma contagem do número de observações. Na base de dados *total case trimester*, os tamanhos dos *clusters* são: "51 41 46". O modelo de agrupamento k-means consiste em três vetores dos valores médios para cada uma das variáveis. Para avaliar a qualidade do modelo, pode-se utilizar as médias seguidas por uma medida simples, calculando o valor do *Within cluster sum of squares* para os três *clusters*: 3.517266 - 2.713257 - 2.431801. A medida utilizada é a soma dos quadrados das diferenças entre as observações (instâncias) dentro de cada um dos três *clusters*.

Em seguida, serão apresentados os números de *centróides* para cada *cluster* entre as variáveis medidos pela distância euclidiana. Os *clusters* com valores mais altos em um determinado atributo indicam que as instâncias neste *cluster* tendem a ter valores mais altos nesse atributo. Por outro lado, os *clusters* com valores mais baixos em um determinado atributo indicam que as instâncias nesse *cluster* tendem a ter valores mais baixos nesse atributo. Pode-se observar esses valores das somas dos *centróides* na Figura 25.

Figura 25 – Cálculos dos *Cluster Centróides - Total de Case Trimestre*

```
Cluster centers:

  population median_age life_expectancy human_development_index
1 0.03992689  0.4187402       0.6923323                0.6256779
2 0.01712776  0.7868978       0.8482659                0.8454732
3 0.01547933  0.1329734       0.3188524                0.2682721
  period_total_case
1       0.037244508
2       0.051258066
3       0.003760051
```

Fonte: Própria Autora

Com o uso do algoritmo de *K-means*, é possível gerar um gráfico de dispersão em que as instâncias são coloridas de acordo com o *cluster* ao qual pertencem, sendo representados por cores distintas, como verde, preto e vermelho. Dessa forma, é possível visualizar como as instâncias de cada *cluster* estão distribuídas em relação às duas variáveis, selecionando a opção *Data* na ferramenta *Rattle*.

Observa-se as instâncias dos *clusters* concentradas em determinadas faixas de valores para as demais variáveis, o que indica um fator importante de definição desses *clusters*. Pode-se ver alguma separação clara entre os *clusters* no eixo horizontal e vertical entre a variável *period total case* e as demais variáveis *population, median age, life expectancy, human development index*, encontrando assim alguns padrões observados se há uma relação entre a incidência de casos da COVID-19 entre as variáveis dos indicadores socioeconômicos e ambientais. Também há a presença de alguns *outliers* nos grupos, então têm-se as seguintes observações:

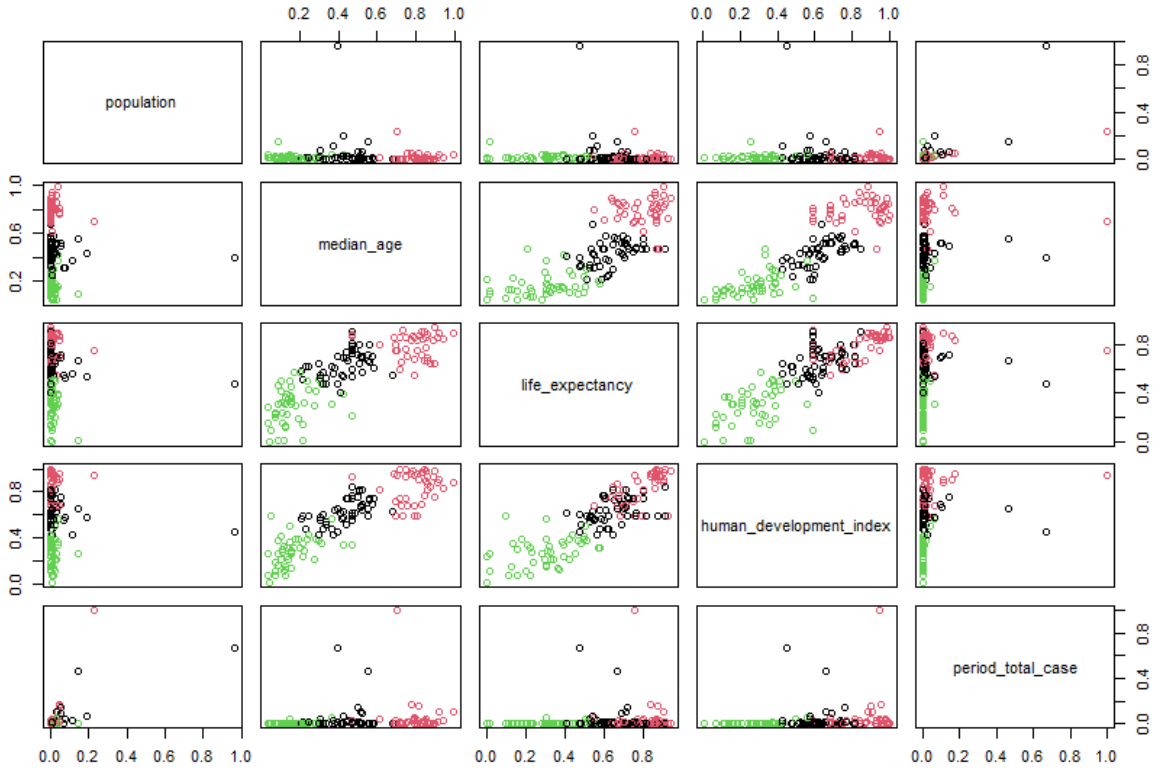
Cluster 1: Pode conter países com um alto número de casos por COVID-19, uma grande população, uma média de idade intermediária a alta, uma expectativa de vida alta e um IDH alto. Isso pode indicar que esses países têm uma população relativamente protegida contra a COVID-19 devido a um sistema de saúde e desenvolvimento humano robusto, mas ainda estão enfrentando desafios significativos em relação à transmissão do vírus devido a uma grande população. Este *cluster* pode ser indicativo de áreas que precisam de mais recursos para conter a transmissão do vírus.

Cluster 2: Pode conter países com um número moderado de casos por COVID-19, uma população intermediária, uma média de idade intermediária, uma expectativa de vida intermediária e um IDH intermediário. Isso pode indicar que esses países estão enfrentando desafios moderados em relação à transmissão do vírus, mas têm recursos e sistemas de saúde e desenvolvimento humano adequados para lidar com a situação. Este *cluster* pode

ser indicativo de áreas que precisam continuar a investir em recursos para prevenir a transmissão do vírus.

Cluster 3: Pode conter países com um baixo número de casos por COVID-19, uma população pequena, uma média de idade baixa a intermediária, uma expectativa de vida baixa a intermediária e um IDH baixo a intermediário. Isso pode indicar que esses países têm enfrentado menos desafios em relação à transmissão do vírus, mas têm sistemas de saúde e desenvolvimento humano menos robustos e, portanto, precisam de mais recursos para prevenir a propagação do vírus. Este *cluster*, pode ser indicativo de áreas que precisam de ajuda e recursos adicionais para lidar com a pandemia. Como podemos observar na Figura 26.

Figura 26 – Gráfico de Dispersão dos *Clusters* - *Total Case Tremester*



Fonte: Própria Autora

Após gerar o modelo de *clustering* apresentado na Figura 14, começamos com a contagem do número de observações dentro de cada *cluster*. Na base de dados *total death per million*, temos esses valores: "46 57 35".

Para avaliar a qualidade do modelo, calculamos a medida de qualidade conhecida como *Within cluster sum of squares* para os três *clusters*, seguida pelas médias de cada variável. Os valores obtidos foram: 4.775886, 6.684055 e 4.644913. Essa medida avalia a soma dos quadrados das diferenças entre as observações (instâncias) dentro de cada um

dos três *clusters*.

Em seguida, apresentamos os números dos *centróides* para cada *cluster*, conforme ilustrado na Figura 27.

Figura 27 – Cálculos dos *Cluster Centróides - Total Death Per Million*

```
Cluster centers:

  median_age extreme_poverty cardiovasc_death_rate life_expectancy
1 0.37180757      0.13195828             0.4661851       0.6031508
2 0.69095729      0.08748457             0.1902935       0.8342042
3 0.09278912      0.41681768             0.4527351       0.2702995
  period_deaths_per_million
1                 0.1164782
2                 0.2973766
3                 0.0244204
```

Fonte: Própria Autora

O gráfico de dispersão plotado no *K-means* permite visualizar a distribuição das instâncias de cada *cluster* em relação às duas variáveis em questão. Através da função *Data*, é possível perceber que as instâncias dos *clusters* estão distribuídas em faixas de valores crescentes em relação às demais variáveis, o que indica um fator importante na definição desses *clusters*. Além disso, é possível observar uma certa separação entre os *clusters* no eixo horizontal e vertical, com relação às variáveis *period death per million*, *median age*, *extreme poverty*, *cardiovascular death rate* e *life expectancy*. Também é possível notar a presença de alguns *outliers* a algumas sobreposições, encontrando assim alguns padrões observados entre as variáveis. Esses padrões podem ser interpretados da seguinte forma:

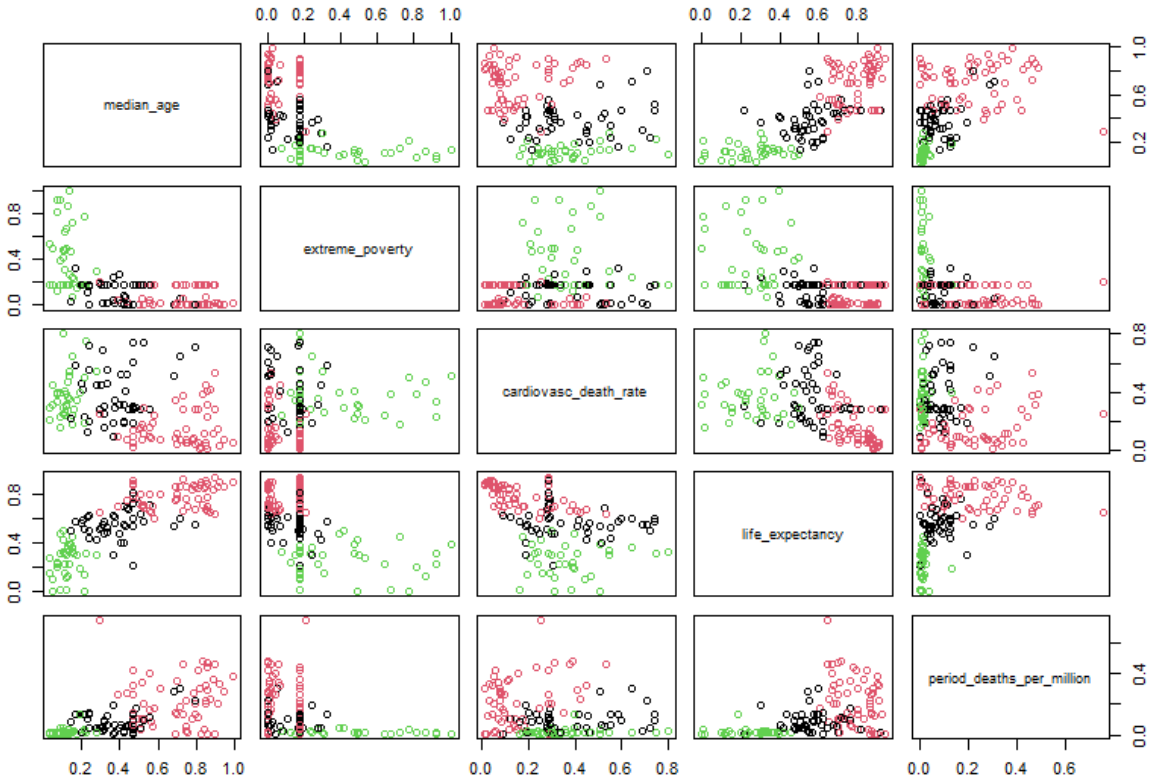
Cluster 1: Pode conter países com uma alta média de idade, alta taxa de mortes cardiovasculares, baixa expectativa de vida e uma proporção significativa de pessoas vivendo em extrema pobreza. É possível que essas países tenham um número elevado de mortes por COVID-19, devido à alta vulnerabilidade da população idosa e doentes cardiovasculares. Este *cluster* pode ser indicativo de áreas que precisam de intervenções de saúde específicas para proteger os grupos vulneráveis.

Cluster 2: Pode conter países com uma baixa média de idade, baixa taxa de mortes cardiovasculares, alta expectativa de vida e uma baixa proporção de pessoas vivendo em extrema pobreza. É possível que essas regiões tenham um número relativamente baixo de mortes por COVID-19, devido à população mais jovem e saudável, que é menos vulnerável à doença. Este *cluster* pode ser indicativo de áreas que precisam manter medidas de prevenção para evitar que o número de casos e mortes por COVID-19 aumente.

Cluster 3: Pode conter países com características intermediárias em relação às

variáveis estudadas, com uma média de idade e expectativa de vida moderadas, uma taxa de mortes cardiovasculares intermediária e uma proporção moderada de pessoas vivendo em extrema pobreza. É possível que essas regiões tenham um número moderado de mortes por Covid-19, mas não tão alto quanto as regiões no *cluster* 1. Este *cluster* pode ser indicativo de áreas que precisam de medidas de prevenção e intervenções de saúde para evitar que o número de casos e mortes por COVID-19 aumente. Como mostra na Figura 28.

Figura 28 – Gráfico de Dispersão dos *Clusters* - *Total Death per Million*



Fonte: Própria Autora

Após a geração do modelo de *clustering* apresentado na Figura 15, é possível realizar outra análise. A primeira delas é o tamanho de cada *cluster*, que corresponde à contagem do número de observações dentro de cada um deles. Na base de dados *people vaccinated*, esses valores são: "50, 42 e 46".

Para avaliar a qualidade do modelo, são apresentadas as médias, seguidas de uma medida simples da qualidade do modelo, o cálculo de K, que é a soma dos quadrados das diferenças entre as observações (instâncias) dentro de cada um dos três *clusters*. Para os três *clusters* do modelo, as medidas de K são: 2.718460, 1.899692 e 2.406035, respectivamente.

Em seguida, são apresentados os números de *centróides* para cada *cluster*, de acordo com a Figura 29.

Figura 29 – Cálculos dos *Cluster Centers - Centróides de People Vaccinated*

```
Cluster centers:

  median_age life_expectancy human_development_index period_people_vaccinated
1  0.4181128       0.6869170               0.6213757                0.04983514
2  0.7788790       0.8510000               0.8453616                0.02741307
3  0.1329734       0.3188524               0.2682721                0.02516278
```

Fonte: Própria Autora

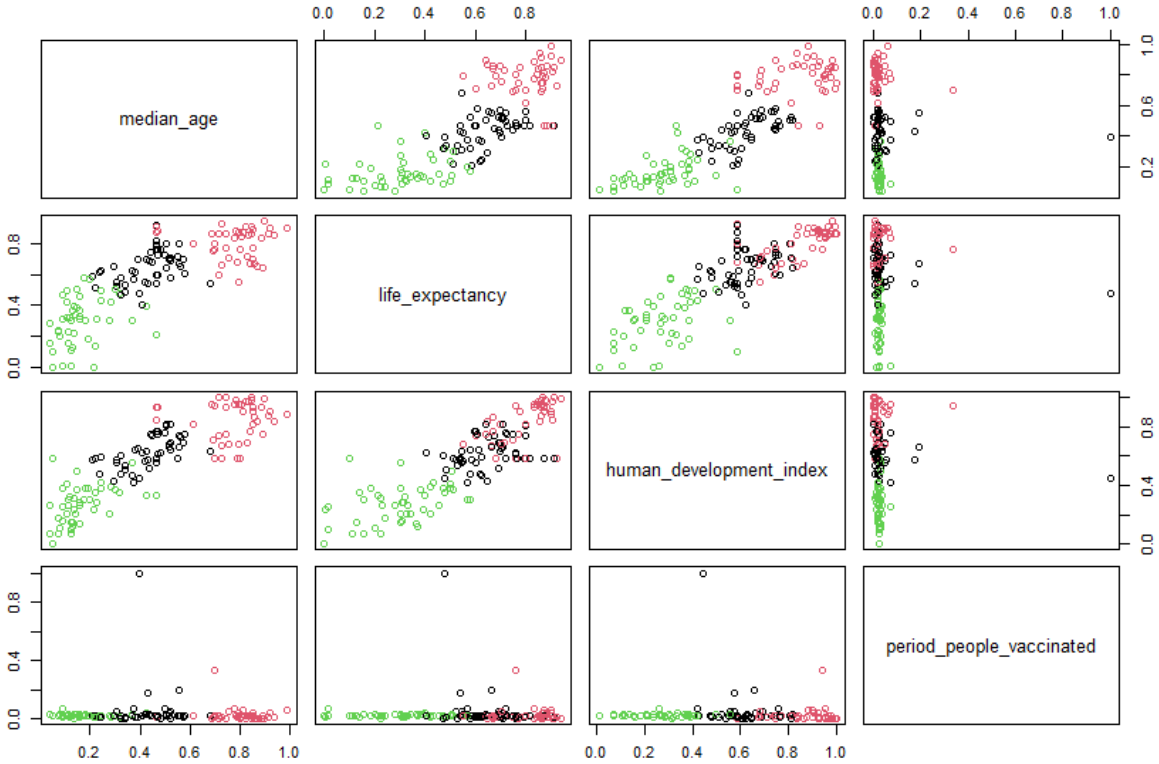
Ao plotar o gráfico de dispersão no *K-means*, é possível visualizar como as instâncias de cada *cluster* estão distribuídas em relação às duas variáveis. Através do botão *Data*, percebe-se que as instâncias dos *clusters* estão concentradas em uma determinada faixa de valores para as demais variáveis, o que indica um fator importante na definição desses *clusters*. Observa-se alguma separação clara entre os *clusters* na horizontal e vertical entre as variáveis *period people vaccinated* e as demais variáveis, como: *media age*, *life expectancy* e *human development index*, bem como a presença de alguns *outliers*. Dessa forma, é possível identificar alguns padrões entre as variáveis:

Cluster 1: Pode conter países com uma alta taxa de pessoas vacinadas por COVID-19, uma média de idade moderada a alta, uma expectativa de vida alta e um IDH alto. Isso pode indicar que esses países têm recursos adequados para a distribuição e administração de vacinas e, como resultado, têm uma população relativamente protegida contra a COVID-19. Este *cluster* pode ser indicativo de áreas que implementaram com sucesso programas de vacinação e investiram em saúde e desenvolvimento humano.

Cluster 2: pode conter países com uma taxa média de pessoas vacinadas por COVID-19, uma média de idade intermediária, uma expectativa de vida intermediária e um IDH intermediário. Isso pode indicar que esses países têm feito esforços para distribuir e administrar vacinas, mas ainda têm um longo caminho a percorrer para atingir a imunidade coletiva. Este *cluster* pode ser indicativo de áreas que precisam de mais recursos para distribuição de vacinas e investimentos em saúde e desenvolvimento humano.

Cluster 3: pode conter países com uma baixa taxa de pessoas vacinadas por COVID-19, uma média de idade baixa a intermediária, uma expectativa de vida baixa a intermediária e um IDH baixo a intermediário. Isso pode indicar que esses países enfrentam desafios significativos em relação à distribuição e administração de vacinas, bem como em relação à saúde e ao desenvolvimento humano em geral. Este *cluster* pode ser indicativo de áreas que precisam de ajuda e recursos adicionais para melhorar a distribuição de vacinas e fortalecer seus sistemas de saúde e desenvolvimento humano. De acordo como mostra na Figura 30.

Figura 30 – Gráfico de Dispersão dos *Clusters* - *Total People Vaccinated*



Fonte: Própria Autora

As avaliações do modelo de *clustering* pode ser observada na aba *evaluate* e com *Score* e *K-means* selecionados, temos as pontuações e identificadores de cada *cluster*, carregado na ordem do conjunto de dados de treinamento e gerando um arquivo no formato CSV.

5.6.3 Regras de Associação

Para compreender e interpretar as análises realizadas, foram utilizadas algumas métricas de avaliação. Para a análise de associação, foram utilizadas duas medidas primárias: suporte e confiança. O suporte mínimo é expresso como uma porcentagem do número total de transações no conjunto de dados, ou seja, é a frequência absoluta com que os itens aparecem juntos em todas as transações. Já a confiança mínima é a probabilidade de que o consequente seja verdadeiro, dado que o antecedente é verdadeiro.

Outra medida de análise é o *lift*, que mede a força da associação entre o antecedente (lhs) e o consequente (rhs). Um valor de *lift* acima de 1 indica que a regra tem mais probabilidade de ser verdadeira do que falsa.

Em relação à análise de avaliação, temos a distribuição das três medidas de acordo com as regras encontradas na Figura 17. Elas atenderam aos critérios de ter um suporte

(*support)* mínimo de 0,1 e uma confiança *(confidence)* mínima de 0,1, configuração padrão do *Rattle*. No entanto, nas regras, o suporte varia de 0,1 a 0,92, a confiança varia de 0,1 a 0,96 e o *lift* de 0,99, ordenadas pelo nível do *lift* na regra. A medida de avaliação *lift* foi a medida escolhida para avaliar e selecionar os padrões de associação mais fortes em todas as avaliações. Ao executar a análise, podemos examinar as regras geradas encontrando os pares de casos com maior suporte e confiança. Combinamos os itens únicos restantes em conjuntos de itens contendo apenas dois itens e retemos apenas aqueles que são frequentes o suficiente.

Em seguida, a regra "casos do 1º trimestre de 2020 -> leitos ocupados"ocorre em 92% das observações no conjunto de dados, o que significa que o suporte da regra é de 92%. Além disso, das observações que contêm casos de COVID-19 no 1º trimestre de 2020, 96% também apresentam leitos ocupados, o que resulta em uma confiança de 96% para a regra. Porém, o *lift* da regra casos do 1º trimestre de 2020 -> leitos ocupados indica que as observações que contêm leitos ocupados são 0.9 vezes menos propensas a ter COVID-19. Um valor de *lift* próximo a 1 indica que não há uma associação forte entre as variáveis. Essas informações são mostradas na Figura 31.

Em resumo, as medidas de suporte, confiança e lift podem ser usadas para avaliar a força das regras de associação em dados relacionados ao COVID-19 e à ocupação de leitos hospitalares. Com base nessas medidas, é possível identificar as associações mais fortes e, assim, ajudar a orientar políticas públicas e alocar recursos hospitalares de maneira mais eficiente.

Figura 31 – Medidas das variáveis Período Total de Casos e Leitos Hospitalares

support	confidence	lift	count
0.9275362	0.9696970	0.9986431	128
0.9275362	0.9552239	0.9986431	128

Fonte: Própria Autora

A análise de avaliação referente às regras encontradas na Figura 18 foi realizada utilizando as medidas de configuração padrão do *Rattle*. As regras seguem: o suporte varia de 0.1 a 0.86, a confiança varia de 0.1 a 0.95 e o *lift* de 1.026, ordenadas pelo nível do *lift* na regra. A medida de avaliação *lift* foi a escolhida para avaliar e selecionar os padrões de associação mais fortes.

Essa regra indica que 86% dos registros apresentam tanto mortes no período do 1º trimestre de 2020 por COVID-19, quanto alto Índice de Gini. A confiança de 95% indica

que, quando um registro apresenta alto valor em mortes por COVID-19, há uma chance de 95% de que o Índice de Gini também seja alto. O valor de *lift* acima de 1 indica que há uma associação positiva entre as variáveis, ou seja, a ocorrência de mortes por COVID-19 aumenta a probabilidade de alto Índice de Gini.

Essa regra indica que em países com alto Índice de Gini, a probabilidade de haver muitas mortes por COVID-19 no 1º trimestre de 2020 é alta. Isso pode ser interpretado como uma possível relação entre desigualdade socioeconômica e maior vulnerabilidade à COVID-19, uma vez que regiões com maior desigualdade socioeconômica tendem a ter menor acesso a recursos de saúde e menor capacidade de adotar medidas de prevenção, como podemos observar na Figura 32.

Figura 32 – Medidas das variáveis Período Total de Mortes e Índice de GINI

suppot	confidence	lift	count
0.8623188	0.9520000	1.026375	119
0.8623188	0.9296875	1.026375	119

Fonte: Própria Autora

A análise de avaliação referente às regras encontradas na Figura 19, utilizando as medidas de configuração padrão do *Rattle*, segue da seguinte forma: o suporte varia de 0.1 a 0.98, a confiança varia de 0.1 a 0.98 e o *lift* varia de 0.99 a 1.17, ordenadas pelo nível do *lift* na regra. A medida de avaliação *lift* foi escolhida para avaliar e selecionar os padrões de associação mais fortes.

A primeira regra, relacionando pessoas vacinadas no 1º trimestre de 2021 e PIB, indica que os países com um PIB per capita mais alto tendem a ter uma taxa de vacinação mais alta. O suporte de 0.96 indica que essa regra se aplica a 96% das observações do conjunto de dados. A confiança de 0.98 indica que, quando o PIB per capita é alto, a taxa de vacinação é alta em 98% dos casos. A medida do *lift* de 0.99 indica que a probabilidade de termos pessoas vacinadas contra a COVID-19 e um PIB alto é 99%. Esses dados sugerem uma forte associação entre o número de pessoas vacinadas contra a COVID-19 no 1º trimestre de 2021 e o PIB, o que pode indicar que países com economias mais fortes foram capazes de investir mais em campanhas de vacinação.

Portanto, as regras encontradas na análise sugerem que fatores como PIB tem forte relação com a taxa de vacinação contra a COVID-19 nos países. É importante ressaltar que esses resultados não devem ser interpretados de forma causal, mas sim como indicativos de possíveis correlações entre as variáveis analisadas. A Figura 33 que apresenta mais detalhes sobre as medidas das regras encontradas.

Figura 33 – Medidas das variáveis Período Pessoas Vacinadas e PIB

support	confidence	lift	count
0.9637681	0.9851852	0.9996732	133
0.9637681	0.9779412	0.9996732	133

Fonte: Própria Autora

Esta análise de avaliação refere-se às regras encontradas na Figura 20 usando as medidas de configuração padrão do *Rattle*. As nossas regras possuem suporte variando de 0.1 a 0.31, confiança variando de 0.1 a 0.93 e *lift* de 1.51, ordenadas pelo nível do *lift* na regra. A medida de avaliação *lift* foi escolhida para avaliar e selecionar os padrões de associação mais fortes.

Isso indica que a regra de associação é verdadeira em 93% dos casos. Com suporte 0.31, indica que 31% dos países no conjunto de dados têm baixa expectativa de vida e alta taxa de mortalidade por COVID-19 no 1º trimestre de 2020. Além disso, a regra de associação tem um valor de *lift* acima de 1, o que indica que países com baixa expectativa de vida têm 1,51 vezes mais chances de alta taxa de mortalidade, ou seja, estão positivamente associados com mortes por milhão de pessoas por COVID-19 no 1º trimestre de 2020. Isso pode ser observado na Figura 34.

Figura 34 – Medidas das variáveis Período de Mortes por Milhões de Pessoas e Expectativa de Vida

support	confidence	lift	count
0.3115942	0.9347826	1.5176471	43
0.3115942	0.5058824	1.5176471	43
0.1159420	1.0000000	1.5000000	16
0.1159420	0.1739130	1.5000000	16
0.1739130	0.8888889	1.3333333	24
0.1739130	0.2608696	1.3333333	24
0.3043478	0.4941176	0.7411765	42
0.3043478	0.4565217	0.7411765	42

Fonte: Própria Autora

Dado o conjunto de dados de casos de COVID-19 por milhão de pessoas no 4º trimestre de 2020, com relação a média de idade baixa, foram obtidas regras de associação com as seguintes medidas de suporte, confiança e *lift*: Suporte = 0.12, Confiança = 0.89, *Lift* = 2.32

Isso indica que, para as transações que possuem média de idade baixa, em 12% delas há alta ocorrência de casos de COVID-19 por milhão de pessoas. Além disso, a associação entre média de idade baixa e casos de COVID-19 por milhão de pessoas é 2.32 vezes mais forte do que se essas variáveis fossem independentes.

Já na relação entre a idade média da população e o número de casos de COVID-19 por milhão de pessoas no 1º e 2º trimestre de 2020, com um suporte de 42%, confiança de 84% e lift de 1,36, podemos inferir que a idade média alta tem uma forte associação positiva com a incidência de casos de COVID-19 por milhão de pessoas nos períodos dos 1º e 2º trimestre de 2020.

Essa informação pode ser útil para entender melhor como a idade média da população afeta os casos de COVID-19 por milhão de pessoas em um determinado trimestre e pode ser usada para direcionar políticas de saúde pública no sentido de priorizar ações de prevenção e tratamento em regiões com população de idade média durante surtos de COVID-19. Além disso, essas informações podem ser úteis para alocar recursos de saúde e insumos de acordo com as necessidades da população afetada. De acordo com o que podemos observar na Figura 35.

Figura 35 – Medidas das variáveis Período de Casos por Milhões de Pessoas e Idade Média

support	confidence	lift	count
0.1231884	0.8947368	2.329692	17
0.1231884	0.3207547	2.329692	17
0.4275362	0.8428571	1.368403	59
0.4275362	0.6941176	1.368403	59
0.1014493	0.1647059	1.082353	14
0.1014493	0.6666667	1.082353	14

Fonte: Própria Autora

6 Discussão dos Resultados

Neste capítulo, são apresentados os resultados obtidos com a execução da metodologia proposta, por meio da aplicação do KDD nas bases de dados da COVID-19 e dos indicadores socioeconômicos e ambientais. Na Seção 5.1, é feita a contextualização da validação dos resultados. Na Seção 5.2, são mostradas as etapas de preparação do ambiente que foram realizadas durante a obtenção dos resultados. Na Seção 5.3, são apresentados os resultados obtidos por meio da RSL, e na Seção 5.4, são apresentados os resultados obtidos na etapa do KDD, destacando a etapa de interpretação e avaliação dos dados.

6.1 Contextualização

Para a execução da metodologia proposta, foi constatada, durante a etapa inicial de coleta de dados nos repositórios, a necessidade de selecionar uma amostra de 198 registros para cada país, realizados por trimestre, no período de 2020 a 2021, gerando assim 5 conjuntos de dados com informações sobre o total de casos, total de mortes, total de casos e mortes por milhão de pessoas e pessoas vacinadas contra a COVID-19, seguidos pela extração de dados dos Indicadores Socioeconômicos e Ambientais. Os atributos necessários para o processo de mineração de dados estão destacados na Tabela 4. Após a extração das bases de dados, foi possível realizar as cincos etapas do KDD.

6.2 Preparação do Ambiente

Os experimentos foram realizados em um computador com processador Intel(R) Core(TM) I5-3210M CPU 2.50GHz, 16 GB de RAM DDR3 e adaptadores de vídeo Intel(R) HD G*Graphics* 4000, executando no sistema operacional *Windows 10 Pro 64 bits*.

Todo o processo de extração de conhecimento foi desenvolvido em um notebook utilizando o ambiente *Jupyter Notebook* e a linguagem de programação *Python*, juntamente com bibliotecas como *numpy, pandas, seaborn, matplotlib e scikit-learn*. Também foi utilizado o software *RStudio* e o pacote *Rattle* da linguagem R.

6.3 Resultados da RSL

Os resultados obtidos na RSL forneceram dados importantes para o estudo da pandemia da COVID-19 e sua relação com indicadores socioeconômicos e ambientais. Foi possível identificar correlações entre perfis socioeconômicos e indicadores climáticos e de poluição do ar com o aumento do número de mortes por COVID-19 e impactos no PIB e

taxa de desemprego. Além disso, o IDH foi identificado como um fator significativo para doenças infecciosas emergentes e consequências na saúde pública.

Em relação às técnicas de *Machine Learning* mais utilizadas, SVM, MLP, RF e RNA foram identificados como modelos que apresentaram resultados satisfatórios na análise da COVID-19. Quanto aos softwares de análise de dados mais utilizados, o *Oryx Apache* e *Python* com a biblioteca *scikit-learn* foram os mais citados.

Esses resultados podem ser úteis para o desenvolvimento de políticas públicas e estratégias para o combate à pandemia da COVID-19, além de fornecer informações para pesquisas futuras nessa área.

6.4 Resultados da Etapa do KDD

Como resultado do processo do KDD foi possível identificar alguns *insights* na etapa de avaliação de desempenho dos modelos através do método de Mineração de Dados descritiva por meio da Correlação, Agrupamento (Clustering) e Regra de Associação. A mineração de dados trata da construção de modelos que nos dão *insights* sobre o mundo e como ele funciona. Mas ainda mais do que isso, esses modelos costumam ser úteis para nos orientar sobre como lidar e interagir com o mundo real.

6.4.1 Resultados das Correlações

O resultado da correlação através do gráfico de dispersão das variáveis *total case 1st trimester 2020* e GDP US$, GNI S$ com correlação de *Pearson* de 0.82 sugere uma forte correlação positiva entre as variáveis. Isso significa que à medida que o Produto Interno Bruto (GDP) e a Renda Nacional Bruta (GNI) aumenta, o número total de casos de COVID-19 tendem a aumentar também. Ou seja, países com economias maiores tendem a ter um maior número de casos de COVID-19. Essa correlação pode ter várias explicações, incluindo o fato de que países mais ricos tendem a ter uma população mais densamente concentrada e maiores oportunidades de viagens internacionais, o que pode aumentar a transmissão do vírus. Além disso, países mais ricos também podem ter mais recursos para testar e identificar casos de COVID-19, o que pode contribuir para o aumento no número total de casos.

Na correlação de *Spearman* obteve-se como resultado a relação entre duas variáveis, neste caso, o número *total case per million 4th trimester 2021* e as variáveis *median age* (idade média da população) e *human development index*. A correlação de *Spearman* de 0,61% com a variável *median age* indica que países com uma média de idade mais alta tendem a ter um maior número de casos de COVID-19 por milhão de pessoas. Isso pode ser explicado pelo fato de que pessoas mais velhas são mais vulneráveis à doença e, portanto, os países com uma população mais velha podem ter mais casos. No caso do

IDH, a correlação de *Spearman* mostrou uma tendência positiva entre as variáveis *total case per million 4th trimester 2021* e *human development index*. Ou seja, países com um índice de desenvolvimento humano mais alto tendem a ter um maior número de casos de COVID-19 por milhão de pessoas no 4º trimestre de 2021. Por outro lado, a correlação de *Spearman* de 0,08% entre o número total de casos de COVID-19 por milhão de pessoas no 3º trimestre de 2020 e a população com 65 anos ou mais sugere uma correlação positiva muito fraca, indicando que não há uma relação clara entre o número total de casos de COVID-19 neste trimestre e a proporção de idosos na população.

Essa informação é importante para entendermos como diferentes fatores demográficos e socioeconômicos podem influenciar a disseminação do vírus em diferentes países. Por exemplo, países com uma população mais envelhecida podem precisar de estratégias específicas para proteger as pessoas idosas e reduzir a disseminação do vírus em lares de idosos. Da mesma forma, países com um IDH mais baixo podem precisar de mais recursos e apoio para implementar medidas de prevenção e controle da COVID-19.

O gráfico de dispersão das variáveis *hospital beds per thousand* e *period people vaccinated* sugere uma possível relação fraca ou inexistente entre as variáveis, uma vez que não há um padrão claro de relação entre elas. Esse resultado é consistente com a correlação de *Kendall* de -0,2%, que indica uma correlação fraca e negativa entre as duas variáveis. O que ocorre com a correlação de *Kendall* de 0,50%, que indica uma correlação positiva moderada entre as variáveis *period people vaccinated* e *population*. Isso significa que, à medida que a população de uma determinada região aumenta, há uma tendência de que o número de pessoas vacinadas também aumente. Para uma compreensão melhor desses resultados, apresentamos a Tabela 6 com os melhores resultados de correlações entre as variáveis.

Tabela 6 – Tabela dos Melhores Resultados das Correlações

Tipo	Variáveis		Magnitude	Correlações
Pearson	total com 1ºst trimester 2020	GPD / GNI	0.82%	Forte
Spearman	total case per million 4st trimester 2021	median age / human development index	0.61%	Forte
Kendall	people vaccinated	population	0.50%	Moderada

Como destaque dos resultados obtidos temos o melhor padrão encontrado na correlação de *Pearson* através do gráfico de dispersão das variáveis, é possível visualizar a relação entre elas, e observar a tendência linear crescente, o que reforça a forte correlação entre as variáveis total de casos do primeiro trimestre de 2020 e o PIB e o GNI são bastante significativos. A forte correlação positiva de 0.82 indica que há uma relação muito forte entre essas variáveis, sugerindo que países com um PIB e GNI mais altos podem estar mais propensos a um maior número de casos da COVID-19, é importante para entendermos a dinâmica da pandemia em diferentes países, e pode ajudar na formulação de políticas públicas mais efetivas para o controle da doença.

6.4.2 Resultados do Agrupamento

Os resultados referentes ao agrupamento obtidos através do gráfico de dispersão do *k-means* com três *clusters* mostram como os países podem ser agrupados com base em seu número de casos de COVID-19, população, idade média, expectativa de vida e IDH, e como esses *clusters* podem estar relacionados com a eficácia da resposta à pandemia e o bem-estar geral da população. Isso pode ajudar a identificar padrões e áreas específicas que precisam de ajuda e recursos adicionais para prevenir a propagação do vírus e fortalecer seus sistemas de saúde e desenvolvimento humano.

A outra análise dos resultados mostra que o gráfico de dispersão do *k-means* com três *clusters* demonstra como os países podem ser agrupados com base em suas características demográficas, socioeconômicas e de saúde, e como esses *clusters* podem ser relacionados com o número de casos de mortes por COVID-19 por milhão de pessoas entre idade média, extrema pobreza, taxa de morte cardiovascular e expectativa de vida, e como esses *clusters* podem estar relacionados. Isso pode ajudar a identificar padrões e áreas específicas que precisam de medidas de prevenção e intervenções de saúde para combater a pandemia.

E por fim, o resultado se resume em um gráfico de dispersão do *k-means* com três *clusters* que apresenta como os países podem ser agrupados com base em sua taxa de pessoas vacinadas por COVID-19, idade média, expectativa de vida e IDH, e como esses *clusters* podem ser relacionados com a eficácia da distribuição de vacinas e o bem-estar geral da população. Isso pode ajudar a identificar padrões e áreas específicas que precisam de ajuda e recursos adicionais para melhorar a distribuição de vacinas e fortalecer seus sistemas de saúde e desenvolvimento humano.

Podemos perceber que os resultados dos gráficos de dispersão dos *clusters* mostraram maiores relações entre as variáveis *life expectancy, human development e median age*, que mais se repetiram e obtive bons resultados na identificação de grupos de objetos similares. Para compreensão dos resultados obtidos será mostrado os melhores resultados dos agrupamentos na Tabela 7.

Tabela 7 – Tabela dos Resultados dos Agrupamentos

Gráfico de Dispersão do K-Means			
Dados da COVID-19	**Dados dos Indicadores**	**Clusters**	**Resultados**
Total Case Trimester	population, media age, life expectancy e IDH	3	Identificar padrões e áreas específicas que precisam de ajuda e recursos adicionais para prevenir a propagação do vírus e fortalecer seus sistemas de saúde e desenvolvimento humano.
Total Deaths per Million	median age, extreme poverty, cardiovac death rate, life expectancy	3	Pode ajudar a identificar padrões e áreas específicas que precisam de medidas de prevenção e intervenções de saúde para combater a pandemia
People Vaccinated	media age, life expectancy e IDH	3	Pode ajudar a identificar padrões e áreas específicas que precisam de ajuda e recursos adicionais para melhorar a distribuição de vacinas e fortalecer seus sistemas de saúde e desenvolvimento humano

Como destaque dos melhores resultados dos gráficos de dispersão dos *clusters* são importantes porque eles ajudam a identificar padrões e relações entre as variáveis. No caso específico mencionado, podemos perceber que as variáveis *life expectancy, human*

development e median age apresentaram uma maior correlação entre os dados da COVID-19 e também se relacionaram bem com outros atributos do conjunto de dados, o que indica que essas variáveis são importantes para entender a dinâmica da COVID-19 em diferentes países.

Ao identificar grupos de objetos similares, podemos entender melhor como a COVID-19 afetou diferentes países de maneiras semelhantes ou distintas, o que pode ser útil para orientar políticas públicas e medidas de saúde para combater a pandemia. Além disso, a identificação desses grupos pode indicar tendências em relação à evolução da doença em diferentes países, permitindo uma melhor compreensão das dinâmicas epidemiológicas e aprimorando a previsão de surtos futuros.

6.4.3 Resultados da Regra de Associação

Com base nos resultados das regras de associações descobertas, pode-se concluir que há uma forte relação entre o número de casos de COVID-19 no 1º trimestre de 2020 e o número de leitos hospitalares ocupados. Isso pode ser útil para os gestores de saúde pública, que podem usar essas informações para planejar e gerenciar melhor o atendimento hospitalar durante a pandemia de COVID-19. Por exemplo, se os casos de COVID-19 estiverem aumentando, os gestores de saúde podem se preparar melhor para lidar com um possível aumento no número de leitos hospitalares ocupados.

A regra de associação das variáveis sugere que a ocorrência de mortes no período do 1º trimestre de 2020 COVID-19 e um alto índice de Gini estão ligeiramente mais associados do que seria esperado ao acaso, mas essa associação não é muito forte. Portanto, pode-se dizer que há uma tendência de que um alto índice de Gini esteja associado com um maior número de mortes no período do 1º trimestre de 2020 COVID-19.

O resultado da regra de associação indica que há uma forte relação entre a variável pessoas vacinadas por COVID-19 no 1º trimestre de 2021 e a variável PIB alto. Esses resultados sugerem que há uma forte relação entre o nível de desenvolvimento econômico de um país (medido pelo PIB) e sua capacidade de vacinar sua população contra a COVID-19. Países com um alto PIB podem ter mais recursos para investir na produção e distribuição de vacinas, bem como na implementação de programas de vacinação.

O resultado dessa regra indica que há uma relação entre as variáveis casos de mortes por COVID-19 no 1º trimestre de 2020 e expectativa de Vida baixa. Essa regra pode ser analisada como uma indicação de que locais com baixa expectativa de vida podem estar enfrentando um impacto mais severo da COVID-19. Isso pode ser devido a uma série de fatores, como desigualdade social, falta de acesso a serviços de saúde adequados e estilo de vida menos saudável.

O resultado dessa regra indica que há uma forte relação entre a variável idade

média baixa e a variável casos de COVID-19 por milhão de pessoas. Esses resultados sugerem que países com populações mais jovens podem ter uma maior incidência de casos de COVID-19 por milhão de pessoas no 4º trimestre de 2020. Isso pode ser explicado pelo fato de que pessoas mais jovens são geralmente mais ativas e mais propensas a se socializarem em grupos, o que aumenta o risco de contrair e disseminar o vírus. Para compreensão dos resultados obtidos será mostrado os melhores resultados dessas regras na Tabela 8.

Tabela 8 – Tabela dos Melhores Resultados das Regras de Associações

Antecedente(lhs)	**Consequente (rhs)**	**Ordenada (Lift)**	**Suporte**	**Confiança**
periodo total casos = casos 1º trimestre 2020	leitos hospitares=ocupados	0.99	0.92	0.96
priodo total de mortes=mortes 1º trimestre 2020	indice GINI=alto	1.02	0.86	0.95
priodo mortes por mihão=morte por milhão 1º trimestte 2021	expectativa de vida=baixa	1.51	0.31	0.93
periodo casos milhão=casos por milhão 4º trimestre 2021	idade médida=baixo	2.32	0.12	0.89
periodo pessoa vacinada=pessoa vacinada 1º trimestre 2021	PIB=alto	0.99	0.96	0.98

Para destacar o melhor resultado que foi obtido pela Regra de Associação, temos a regra com *lift* de 1.02, suporte de 0.86 e confiança de 0.95 entre o antecedente período total de mortes=mortes 1º trimestre 2020 e o consequente indice GINI=alto é bastante interessante e pode fornecer *insights* importantes para os pesquisadores e tomadores de decisão. O *lift* de 1.02 indica que a relação entre as duas variáveis não é forte, porém relevante, mas ainda assim, há uma tendência de ocorrência conjunta entre elas. O suporte de 0.86 indica que a regra de associação é válida para uma grande parte do conjunto de dados analisado, enquanto a confiança de 0.95 indica que há uma alta probabilidade de que a ocorrência do consequente indice GINI=alto seja causada pela ocorrência do antecedente periodo total de mortes=mortes 1º trimestre 2020.

Dessa forma, pode-se inferir que países com um alto índice GINI (indicando uma grande desigualdade socioeconômica) podem ter uma tendência a sofrer mais com a COVID-19 durante o primeiro trimestre de 2020, quando comparados a países com um índice GINI mais baixo. Essa informação pode ser útil para guiar políticas públicas e estratégias de prevenção e combate à pandemia.

Em resumo, a utilização do KDD permite a análise de dados de forma automatizada. Essas técnicas permitem a identificação de padrões e relações em dados complexos, possibilitando a tomada de decisões e a geração de *insights* em diversas áreas, incluindo a análise de indicadores socioeconômicos e ambientais em relação à COVID-19.

6.5 Publicação do Artigo Científico

A princípio, por meio da RSL, foi possível identificar vários indicadores socioeconômicos relacionados à COVID-19, como concentração de poluentes (PM_{10}, $PM_{2,5}$ e NO_2), bem como os algoritmos de aprendizado de máquina mais utilizados para análise de

dados da COVID-19 na literatura, tais como SVM, MLP, RF, RNA e Regressão Linear. Com base nessa pesquisa, um artigo foi submetido para publicação no evento XVIII Simpósio Brasileiro de Sistemas de Informação (SBSI 2022) de Qualis B2, que teve como Trilha de Temas Emergentes a Transformação Digital e Problemas Socio-Urbanos, e foi organizado pela Universidade Tecnológica Federal do Paraná, na cidade de Curitiba - PR, entre os dias 16 e 19 de maio de 2022. O artigo foi aceito e publicado na biblioteca digital SBC OpenLib (SOL) com acesso aberto, conforme pode ser acessado no link https://sol.sbc.org.br/index.php/sbsi_estendido/article/view/21602/21426.

7 Conclusão

Neste trabalho, a revisão bibliográfica foi conduzida por meio de uma RSL. Essa abordagem permitiu identificar uma ampla gama de estudos que investigaram a relação entre a COVID-19 e os Indicadores Socioeconômicos e Ambientais. Para atingir os objetivos específicos propostos, a experimentação da mineração proposta no Capítulo 5 foi aplicada, seguindo as etapas do KDD. Isso permitiu demonstrar, na prática, os passos para a descoberta do conhecimento em bases de dados da COVID-19 e de indicadores socioeconômicos e ambientais. Sendo assim, considera-se que o objetivo geral desta dissertação foi alcançado.

Inicialmente, realizou-se um estudo de revisão de literatura, que enriqueceu o tema proposto e permitiu a implementação do processo de KDD para a descoberta de conhecimento em bases de dados. A execução de todas as etapas da mineração de dados possibilitou explorar e analisar, de forma eficaz e eficiente, os dados da COVID-19 e dos indicadores socioeconômicos e ambientais, produzindo resultados satisfatórios quanto à problemática dos fatores econômicos globais.

Embora o uso do KDD tenha sido fundamental para a descoberta de padrões e *insights* valiosos em nossa análise de dados sobre a COVID-19 e indicadores socioeconômicos e ambientais, é importante destacar suas limitações. Uma delas é a falta de dados confiáveis e atualizados em muitos países, o que pode afetar a precisão das análises. Portanto, é importante realizar análises mais aprofundadas e considerar as limitações mencionadas para evitar conclusões equivocadas sobre a relação entre a COVID-19 e os indicadores socioeconômicos e ambientais.

A análise e discussão dos resultados foi capaz de descobrir as desigualdades socioeconômicas expostas pela pandemia em todo o mundo. Populações mais pobres e vulneráveis foram mais afetadas pela doença, devido à falta de acesso a cuidados de saúde adequados e à necessidade de continuar trabalhando para sobreviver. Essas disparidades mostraram a importância de políticas públicas que visem reduzir a pobreza e melhorar o acesso a serviços de saúde. A urbanização e densidade populacional também teve influência, onde cidades densamente povoadas foram particularmente afetadas pela pandemia devido à facilidade de transmissão do vírus em espaços confinados e superlotados. As lições aprendidas sugerem que as cidades precisam repensar seus modelos de planejamento urbano para torná-las mais adaptáveis a situações como pandemias futuras.

Como principais contribuições desta dissertação, pode-se citar: Durante o processo de KDD para análise dos dados da COVID-19 e indicadores socioeconômicos e ambientais, foi possível obter contribuições significativas para o entendimento da relação entre esses fatores. Foi possível identificar, por exemplo, a correlação alta positiva entre o PIB e casos

da COVID-19 em alguns países. Além disso, a regra de associação entre casos de COVID-19 e leitos hospitalares ocupados permitiu entender a importância da disponibilidade de recursos médicos para o controle da pandemia. Outra contribuição importante foi a identificação da associação entre o índice de Gini e as mortes por COVID-19 em alguns países, evidenciando a necessidade de se pensar em políticas públicas que combatam a desigualdade social como forma de prevenção e controle da pandemia. Portanto, podemos concluir que o processo de KDD permitiu a identificação de padrões e relações importantes entre os dados analisados, fornecendo contribuições significativas para o entendimento da pandemia de COVID-19 e sua relação com os indicadores socioeconômicos e ambientais. Essas informações podem ser utilizadas para orientar políticas públicas e estratégias de combate à pandemia e seus efeitos sobre a sociedade.

Após a aplicação do KDD em dados relacionados à COVID-19 e indicadores socioeconômicos e ambientais, algumas possibilidades de trabalhos futuros surgem. Uma delas é a utilização de técnicas mais avançadas de aprendizado de máquina, como redes neurais, para modelar e prever o comportamento da doença e suas interações com os indicadores através de série temporais. Outra possibilidade é a realização de estudos mais detalhados e específicos sobre regiões ou países específicos, levando em consideração as características locais. Além disso, é importante considerar a inclusão de novos dados e indicadores, que possam trazer mais informações relevantes para a análise. Por fim, é necessário desenvolver metodologias que permitam a análise em tempo real dos dados, possibilitando uma resposta mais rápida e efetiva em situações de emergência, como a atual pandemia de COVID-19.

Os dados do mundo real geralmente exigem grande esforço em garantir a qualidade dos dados que se coleta, sempre poderá ocorrer erros. Por esse motivo é importante ressaltar que a escolha da extração dos dados da COVID-19 e dos Indicadores Socioeconômicos e Ambientais durante os 4 trimestres de 2020 e de 2021 foi crucial para a compreensão dos impactos da pandemia em diferentes momentos do ano. A análise dos dados em diferentes trimestres permitiu observar as mudanças nos indicadores e a identificação de padrões de como eles podem afetar a disseminação da COVID-19. Além disso, a coleta de dados ao longo do período de dois anos forneceu uma visão mais ampla da evolução da pandemia e permitiu avaliar a possível comparação do impacto no primeiro trimestre, quando a pandemia estava apenas começando, com o impacto no quarto trimestre, quando muitos países já haviam iniciado seus programas de vacinação. Essa escolha de extração de dados pode fornecer informações valiosas para tomadores de decisão que buscam entender os impactos da pandemia em diferentes aspectos da sociedade. Para garantir a qualidade e confiabilidade dos dados coletados, foram adotados critérios de seleção, como a verificação da consistência dos dados e a avaliação da fonte de origem.

Referências

AHMAD, S. N. A. F.; HUMAYUN, M.; NASEEM, S.; KHAN, W. A.; JUNAID, K. Prediction of covid-19 cases using machine learning for effective public health management. *Computers, Materials & Continua*, v. 66, n. 3, p. 2265–2282, 2021. Citado 2 vezes nas páginas 31 e 33.

ALI, A. A.; GHAREB, M. I. Knowledge discovery in health domain using deep neural network algorithms. *Passer Journal of Basic and Applied Sciences*, University of Garmian, v. 4, n. Special issue, p. 107–123, 2022. Citado na página 15.

ALODAT, M. Using deep learning model for adapting and managing covid-19 pandemic crisis. *Procedia Computer Science*, v. 184, p. 558–564, 2021. Citado na página 20.

AMOROSO, N.; CILLI, R.; MAGGIPINTO, T.; MONACO, A.; TANGARO, S.; BELLOTTI, R. Satellite data and machine learning reveal a significant correlation between no2 and covid-19 mortality. *Environmental Research*, v. 204, p. 111970, 2021. ISSN 0013-9351. Citado 2 vezes nas páginas 31 e 33.

CAMARGO, A.; AMARAL, ; SILVA, R.; HEINEN, M.; PEREIRA, F. Mineração de dados eleitorais: descoberta de padrões de candidatos a vereador na região da campanha do rio grande do sul. *Revista Brasileira de Computação Aplicada*, v. 8, p. 64–73, 04 2016. Citado na página 17.

CRODA, J. H. R.; GARCIA, L. P. Resposta imediata da vigilância em saúde à epidemia da covid-19. SciELO Public Health, 2020. Citado na página 21.

DOMINGUES, C. M. A. S. Desafios para a realização da campanha de vacinação contra a covid-19 no brasil. SciELO Brasil, 2021. Citado na página 21.

FAYYAD, U. M.; PIATETSKY-SHAPIRO, G.; SMYTH, P. et al. Knowledge discovery and data mining: Towards a unifying framework. v. 96, p. 82–88, 1996. Citado 4 vezes nas páginas 17, 18, 19 e 20.

FERREIRA, J. C.; ROSA, C. R. M.; STEINER, M. T. A. Knowledge discovery in database e data mining: Uma contribuiÇÃo bibliomÉtrica. XXXVIII ENCONTRO NACIONAL DE ENGENHARIA DE PRODUCAO, 2018. Citado na página 19.

GALVAO, M. C. B.; RICARTE, I. L. M. Revisao sistematica da literatura: Conceituacao, producao e publicacao. *Logeion: Filosofia da Informação*, v. 6, n. 1, p. 57–73, set. 2019. Citado na página 26.

GOLDSCHMIDT, R.; PASSOS, E.; BEZERRA, E. Data mining: Conceitos, técnicas, algoritmos, orientações e aplicações. Elsevier Editora Ltda., 2015. Citado 3 vezes nas páginas 14, 19 e 20.

HERVA, M.; FRANCO, A.; CARRASCO, E. F.; ROCA, E. Review of corporate environmental indicators. *Journal of Cleaner Production*, v. 19, n. 15, p. 1687–1699, 2011. ISSN 0959-6526. Citado na página 22.

HYMAN, M.; MARK, C.; IMTEAJ, A.; GHIAIE, H.; REZAPOUR, S.; SADRI, A. M.; AMINI, M. H. Data analytics to evaluate the impact of infectious disease on economy: Case study of covid-19 pandemic. *Patterns*, v. 2, n. 8, p. 100315, 2021. ISSN 2666-3899. Citado 2 vezes nas páginas 31 e 35.

KAMSU-FOGUEM, B.; RIGAL, F.; MAUGET, F. Mining association rules for the quality improvement of the production process. *Expert systems with applications*, Elsevier, v. 40, n. 4, p. 1034–1045, 2013. Citado na página 19.

KANNAN, S.; SUBBARAM, K.; ALI, S.; KANNAN, H. The role of artificial intelligence and machine learning techniques: Race for covid-19 vaccine. *Archives of Clinical Infectious Diseases*, Kowsar, v. 15, n. 2, 2020. Citado na página 21.

KRAUSE, P. R.; FLEMING, T. R.; PETO, R.; LONGINI, I. M.; FIGUEROA, J. P.; STERNE, J. A.; CRAVIOTO, A.; REES, H.; HIGGINS, J. P.; BOUTRON, I. et al. Considerations in boosting covid-19 vaccine immune responses. *The Lancet*, Elsevier, v. 398, n. 10308, p. 1377–1380, 2021. Citado na página 21.

MAGAZZINO, C.; MELE, M.; MORELLI, G. The relationship between renewable energy and economic growth in a time of covid-19: A machine learning experiment on the brazilian economy. *Sustainability*, v. 13, n. 3, 2021. ISSN 2071-1050. Disponível em: <https://www.mdpi.com/2071-1050/13/3/1285>. Citado 4 vezes nas páginas 14, 21, 31 e 34.

MAGAZZINO, C.; MELE, M.; SARKODIE, S. A. The nexus between covid-19 deaths, air pollution and economic growth in new york state: Evidence from deep machine learning. *Journal of Environmental Management*, v. 286, p. 112241, 2021. ISSN 0301-4797. Citado 2 vezes nas páginas 31 e 34.

MAIMON, O.; ROKACH, L. Introduction to knowledge discovery in databases. Springer US, Boston, MA, p. 1–17, 2005. Citado 2 vezes nas páginas 17 e 18.

MELE, M.; MAGAZZINO, C. Pollution, economic growth, and covid-19 deaths in india: a machine learning evidence. *Environmental Science and Pollution Research*, Springer, v. 28, n. 3, p. 2669–2677, 2021. Citado 2 vezes nas páginas 31 e 34.

MIHOUB, A.; SNOUN, H.; KRICHEN, M.; SALAH, R. B. H.; KAHIA, M. Predicting covid-19 spread level using socio- economic indicators and machine learning techniques. p. 128–133, 2020. Citado 2 vezes nas páginas 31 e 33.

NASSER, F. K.; BEHADILI, S. F. A review of data mining and knowledge discovery approaches for bioinformatics. *Iraqi Journal of Science*, p. 3169–3188, 2022. Citado na página 15.

PAUL, A.; ENGLERT, P.; VARGA, M. Socio-economic disparities and COVID-19 in the USA. IOP Publishing, v. 2, n. 3, p. 035017, jul 2021. Disponível em: <https://doi.org/10.1088/2632-072x/ac0fc7>. Citado 2 vezes nas páginas 31 e 34.

PINHEIRO, C. A. R.; GALATI, M.; SUMMERVILLE, N.; LAMBRECHT, M. Using network analysis and machine learning to identify virus spread trends in covid-19. *Big Data Research*, v. 25, p. 100242, 2021. ISSN 2214-5796. Citado na página 20.

PROGRAMME, U. U. N. D. Covid-19 and human development: Assessing the crisis, envisioning the recovery. *UNDP (United Nations Development Programme)*, 2020. Disponível em: <http://hdr.undp.org/en/hdp-covid>. Citado na página 15.

PUNN, N. S.; SONBHADRA, S. K.; AGARWAL, S. Covid-19 epidemic analysis using machine learning and deep learning algorithms. *MedRxiv*, Cold Spring Harbor Laboratory Press, 2020. Citado na página 21.

RAJA, R.; NAGWANSHI, K. K.; KUMAR, S.; LAXMI, K. R. Data mining and machine learning applications. John Wiley & Sons, 2022. Citado na página 20.

SIDDIQUI, M. K.; MORALES-MENENDEZ, R.; GUPTA, P. K.; IQBAL, H. M.; HUSSAIN, F.; KHATOON, K.; AHMAD, S. Correlation between temperature and covid-19 (suspected, confirmed and death) cases based on machine learning analysis. *Journal of Pure and Applied Microbiology*, 2020. Citado 2 vezes nas páginas 31 e 34.

SINGHAL, N. A review on knowledge discovery from databases. *Electronic Systems and Intelligent Computing: Proceedings of ESIC 2021*, Springer, p. 457–464, 2022. Citado na página 15.

SUWIRYA, I. P.; CANDIASA; MADE, I.; DANTES; RASBEN, G. Evaluation of atm location placement using the k-means clustering in bni denpasar regional office. *Journal of Computer Networks, Architecture and High Performance Computing*, v. 4, n. 2, p. 158–168, Jul. 2022. Citado na página 18.

WONG, C.; MICHALOS, A. Economic and social indicators. *Encyclopedia of Quality of Life and Well-Being Research*, Springer Nature, United States, jan. 2014. Citado na página 22.

WORLD, B. World development indicators. 2015. Disponível em: <https://datacatalog.worldbank.org/search/dataset/0037712/World-Development-Indicators>. Citado na página 22.

ZAREMBA, A.; KIZYS, R.; TZOUVANAS, P.; AHARON, D. Y.; DEMIR, E. The quest for multidimensional financial immunity to the covid-19 pandemic: Evidence from international stock markets. *Journal of International Financial Markets, Institutions and Money*, v. 71, p. 101284, 2021. ISSN 1042-4431. Citado 2 vezes nas páginas 31 e 34.

ZHANG, S.; ZHANG, C.; WU, X. Knowledge discovery in multiple databases. Springer Science & Business Media, 2004. Citado na página 18.

ZHOU, P.; YANG, X.-L.; WANG, X.-G.; HU, B.; ZHANG, L.; ZHANG, W.; SI, H.-R.; ZHU, Y.; LI, B.; HUANG, C.-L. et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. *nature*, Nature Publishing Group, v. 579, n. 7798, p. 270–273, 2020. Citado na página 14.