UNIVERSIDADE FEDERAL DO TOCANTINS
CÂMPUS UNIVERSITÁRIO DE PALMAS
CURSO DE CIÊNCIA DA COMPUTAÇÃO

RENATO LUIZ DE ALMEIDA

MINERAÇÃO DE DADOS PARA ENCONTRAR CARACTERÍSTICAS E SINTOMAS EM COMUM ENTRE OS PACIENTES DE COVID-19 NO BRASIL

PALMAS (TO)
2022

RENATO LUIZ DE ALMEIDA

MINERAÇÃO DE DADOS PARA ENCONTRAR CARACTERÍSTICAS E SINTOMAS EM COMUM ENTRE OS PACIENTES DE COVID-19 NO BRASIL

Trabalho de Conclusão de Curso II apresentado à Universidade Federal do Tocantins para obtenção do título de Bacharel em Ciência da Computação, sob a orientação do(a) Prof.(a)
George Lauro Ribeiro de Brito.

Orientador:
George Lauro Ribeiro de Brito

PALMAS (TO)
2022

Dados Internacionais de Catalogação na Publicação (CIP)
Sistema de Bibliotecas da Universidade Federal do Tocantins

A447m Almeida, Renato Luiz de.

Mineração de dados para encontrar características e sintomas em comum entre os pacientes de covid-19 no Brasil. / Renato Luiz de Almeida. – Palmas, TO, 2022.

47 f.

Artigo de Graduação - Universidade Federal do Tocantins – Câmpus Universitário de Palmas - Curso de Ciências da Computação, 2022.

Orientador: George Lauro Ribeiro de Brito

1. Mineração de Dados. 2. Apriori. 3. Covid-19. 4. Regra de Associação. I. Título

CDD 004

Elaborado pelo sistema de geração automática de ficha catalográfica da UFT com os dados fornecidos pelo(a) autor(a).

RENATO LUIZ DE ALMEIDA

MINERAÇÃO DE DADOS PARA ENCONTRAR CARACTERÍSTICAS E SINTOMAS EM COMUM ENTRE OS PACIENTES DE COVID-19 NO BRASIL

Trabalho de Conclusão de Curso II apresentado à UFT – Universidade Federal do Tocantins – Câmpus Universitário de Palmas, Curso de Ciência da Computação foi avaliado para a obtenção do título de Bacharel e aprovada em sua forma final pelo Orientador e pela Banca Examinadora.

Data de aprovação: 3 / 7 / 2022

Banca Examinadora:

_____

Prof. Dr. Patrick Letouzé Moreira

_____

Prof. Dr. David Nadler Prata

Dedico esse projeto a minha mãe
Maria Eunice de Almeida e ao meu
pai Eustáquio Luiz Neto

AGRADECIMENTOS

Gostaria de agradecer aos meus pais que sempre tiveram do meu lado ao longo desses anos de graduação me dando todo suporte e incetivo para que eu pudesse me formar. Aos meus amigos e ao meu irmão que me ajudaram a chegar até aqui. Agradecer ao meu orientador pelo apoio e a pela dedicação ao me ajudar a desenvolver esse projeto e a banca avaliadora pela disponibilidade, pelas considerações, as correções e os ensinamentos.

RESUMO

No final de 2019 surgia uma perigosa variante do vírus da covid, a SARS-CoV-2, esse que causa a covid-19. Esse novo vírus provocou uma pandemia mundial que se estende até os dias de hoje, ele se espalhou rapidamente por todo o globo provocando a contaminação de grande parte da população. Essa pandemia nos colocou diante de grande numero de doentes e mortos, por esse auto nível de gravidade, surgiu danos profundos em nossa sociedade, sendo alguns deles irreversíveis. A partir de então, nossas vidas mudaram completamente, novos hábitos e costumes foram incluídos no nosso dia-a-dia, como isolamento social, distanciamento, higiene pessoal mais rigorosas, novas formas de trabalho, entre outras coisas.

Nesse cenário quando olhamos para computação enxergamos sua grande importância para o mundo, pois ela desempenha um grande papel no combate e no entendimento do vírus causador da covid-19. Focando na área de ciência de dados, esse trabalho tem como objetivo encontrar características e sintomas em comum entre os pacientes aplicando técnicas de mineração nos dados nos casos de covid-19 de todo o Brasil.

Palavra-chave: Mineração de Dados. Covid-19. Apriori. Pandemia.

ABSTRACT

At the end of 2019, a dangerous variant of the covid virus emerged, SARS-CoV-2, the one that causes covid-19. This new virus caused a worldwide pandemic that extends to the present day, it has spread rapidly across the globe causing the contamination of a large part of the population. This pandemic has put us in front of a large number of sick and dead, due to this level of severity, deep damage has arisen in our society, some of which are irreversible. From then on, our lives changed completely, new habits and customs were included in our daily lives, such as social isolation, distancing, stricter personal hygiene, new ways of working, among other things.

In this scenario, when we look at computing, we see its great importance to the world, as it plays a great role in combating and understanding the virus that causes covid-19. Focusing on the area of data science, this work aims to find common characteristics and symptoms among patients by applying data mining techniques in cases of covid-19 from all over Brazil.

Keywords: Data Mining. Covid-19. Apriori. Pandemic.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

Ao longo da nossa história podemos encontrar vários períodos em que a humanidade enfrentou surtos de doenças infecciosas, esses que diversas vezes foram de nível pandêmico onde a disseminação ultrapassou territórios e se espalhou mundialmente. Com grandes impactos em nossa sociedade, esses eventos refletiram ao decorrer do tempo em todos os nossos modelos políticos, educacionais, econômicos e sociais, para ter uma noção, chegaram a interromper guerras e exterminar populações inteiras (HUREMOVIĆ, 2019). A área científica sempre teve grandes avanços nesses períodos epidêmicos, onde observamos que partes como a medicina conseguiu evoluir desenvolvendo métodos de prevenção, de tratamento e controle, assim como também métodos de imunização para combater essas doenças infecciosas.

Atualmente o mundo vem enfrentando a pandemia de covid-19, causada pelo vírus coronavírus SARS-CoV-2, que teve como epicentro a província Hubei na China. Com rápida contaminação global, no dia 30 de janeiro de 2020 a Organização Mundial da Saúde (OMS) já declarava preocupação e alertava todo o mundo da emergência de saúde global, passado os dias só crescia o número de contaminados e mortos por todo o mundo, então dia 11 março a OMS declarou oficialmente a pandemia de Covid-19 colocando o mundo todo em alerta (PINHEIRO, 2020).

Figura 1 – **Pademia de Covid-19 (Foto: Shutterstock, 2020).**

A covid-19 é uma infecção respiratória aguda de potencial grave com números altos de mortalidade e com elevada taxa de transmissão e distribuição global. Sendo uma doença que se manifesta de formas diferentes em cada indivíduo, geralmente a maioria dos infectados desenvolve a doença em grau leve e se recuperam sem hospitalização e já em outros casos onde ela se manifesta de forma mais grave pode necessitar a hospitalização e até causar o óbito do infectado. Os principais sintomas comuns são febre, tosse, cansaço e perda do paladar e olfato; já os sintomas incomuns são dor de cabeça, diarréia, dor de garganta e olhos irritados ou vermelhos;

agora os sintomas graves são dificuldade de respirar ou falta de ar, dor no peito, dificuldades para falar, confusão mental e dificuldades de locomoção (OMS, 2021). Nos casos mais graves o paciente deverá receber auxílio de aparelhos respiratórios para ajudar na respiração.

No Brasil, assim como muitos lugares do mundo, algumas ações de controle de contaminação foram impostas pelas autoridades governamentais recomendadas pela OMS, são elas o uso obrigatório de máscara, o distanciamento social, a proibição de aglomerações, a correta higienização das superfícies e das mãos, toque de recolher e entre outras. Todas essas medidas foram com o intuito de frear os impactos da covid, diminuir a contaminação, evitar mortes e superlotação dos hospitais (WERNECK; CARVALHO, 2020). Apesar de esforços de parte da população para combater a pandemia a desinformação e a negação fez com que algumas pessoas impedisse que o Brasil conseguisse reduzir os impactos negativos em nossa sociedade, o alto número de mortos, mutações do vírus e o grande número de contaminados foram alguma das consequências mais graves.

A Computação é uma grande aliada nessa luta contra a covid-19, os estudos em diversas partes da computação podem auxiliar na hora de entender o problema ou na hora de criar modelos de soluções. Esse trabalho tem como proposta aplicar técnicas de mineração de dados com a intuito de que a partir dos dados de covid-19 seja possível computacionalmente prever números, analisar o progresso, identificar problemas e achar soluções, tendo como principal foco relacionar as características e sintomas em comum entre os pacientes. A partir da obtenção dos dados será aplicado o processo de KDD (Knowledge Discovery in Databases), que é um processo de descoberta de conhecimento em uma base de dados, com o objetivo de estudar esses dados. O processo de KDD consiste em usar nove passos que irão ajudar a extrair informações importantes e úteis de uma base de dados. Esses passos são: compreender o domínio dos dados, selecionar e criar o conjunto de dados, pré-processar e limpar os dados, transformar os dados, realizar a previsão e descrição, escolher o algoritmo de mineração de dados, utilizar o algoritmo de mineração, avaliar os resultados e, por fim, usar o conhecimento descoberto na base (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996).

2 TRABALHOS RELACIONADOS

Esse tópico tem por objetivo mostrar os trabalhos lidos e estudados para o desenvolvimento desse projeto, esses que serviram de referência para iniciar os estudos. Será abordado de forma direta e resumida, quais os objetivo dos trabalhos, quais os métodos usados e como foram desenvolvidos, os principais resultados atingidos, por quem foram escritos e algumas conclusões.

2.1 Discovering symptom patterns of COVID-19 patients using association rule mining

O artigo "Discovering symptom patterns of COVID-19 patients using association rule mining"foi publicado em abril de 2021 no site National Library of Medicine, esse que é um site oficial do governo dos Estados Unidos que publica estudos nas áreas biomédicas e genômicas. O estudo é voltado para a área da medicina e da computação, sendo seus autores dessas áreas (TANDAN et al., 2021).

O principal objetivo do estudo realizado foi encontrar alguns padrões entre os sintomas e as regra gerais de sintomas, relacionando entre os pacientes com COVID-19 a idade, o sexo, as comorbidades e a mortalidade, os dados utilizados foram pegos online em uma base de dados disponibilizada no site Wolfram Data Repository em maio de 2020. O trabalho foi voltado para a área de mineração de dados, onde aplicaram técnicas de aprendizado de máquina baseadas em regras de associação para identificar sintomas frequentes e determinar padrões em regras encontradas.

Com media de idade de 52 anos, cerca de 1560 pacientes com covid-19 fizeram parte desse estudo, onde os resultados foram que os sintomas mais frequentes eram febre com 67%, tosse com 37%, dor no corpo 11% e pneumonia 11%. Das regras geradas pode se observa que tosse, choque séptico e desconforto respiratório eram sintomas que ocorriam frequentemente juntos. O estudo concluiu que regras associadas se diferenciavam de acordo com a idade e o sexo dos pacientes, e que pacientes com problemas cardiovasculares e com sintomas de tosse, febre e pneumonia foram as vitimas mais graves.

2.2 Leveraging Data Science to Combat COVID-19: A Comprehensive Review

Publicado pelo IEEE em agosto de 2020 na "IEEE Transactions on Artificial Intelligence", o artigo "Leveraging Data Science to Combat COVID-19: A Comprehensive Review"tem como objetivo ajudar a comunidade acadêmica e científica das áreas de Inteligencia Artificial e Ciência de Dados a encontrar e a criar soluções que possam auxiliar a combater a covid-19 através da sintetização de recursos disponíveis sobre o assunto (LATIF et al., 2020).

Afirmando a importância da ciência dos dados no combate a pandemia global de covid-

19, o trabalho trás recursos para os pesquisadores que facilita o primeiro passo para a compreensão e a introdução ao assunto. O estudo apresenta informações de dados, ferramentas e fórmulas disponíveis até o momento, problemas relacionados a covid-19 que seriam interessantes de se usar a ciência de dados para combater, os estudos mais relevantes que existem no meio cientifico sobre a ciência de dados na luta contra a covid-19 e por fim os principais obstáculos enfrentados pelos pesquisadores e cientistas de dados.

3 MINERAÇÃO DE DADOS

Desde de o inicio da computação estudar dados tem sido uma atividade bastante fundamental e com o passar do tempo isso vem se tornando cada vez mais importante e necessário, hoje podemos dizer que dados são um dos recursos computacionais mais valiosos, sendo de extrema importância para qualquer empresa ou instituição. Uma das tecnologias mais utilizadas quando se trata de estudar e explorar dados é a Mineração de Dados.

O termo Mineração de Dados se refere ao conjunto de técnicas utilizadas para descobrir conhecimento explorando base de dados, esse processo busca encontrar alguns padrões, irregularidades e correlações entre os dados dessa base. Sendo que cada caso exige uma estratégia de mineração diferente, cabe então ao cientista de dados escolher a técnica adequada ao seu problema, por isso é necessário conhecer bem o tipo de dado que será trabalhado.

Preparar os dados é uma atividade essencial e muito importante para o processo de mineração, pois a partir da preparação dos dados poderemos inicialmente compreender melhor a natureza dos dados e assim escolher o melhor caminho para se seguir. Com esse intuito, alguns passos são necessários, sendo primeiramente a limpeza dos dados, onde será eliminado a inconsistência na base, como a exclusão de valores errados e incompletos, a substituição por padrões mais consistentes em alguns casos, e até mesmo o agrupamento de alguns valores. Em seguida podemos integrar alguns dados ou até mesmo bases, isso vai depender da necessidade encontrada. Outra etapa importante é a transformação dos dados, que é basicamente converte-los para uma forma que possa suprir seus objetivos estabelecidos. Por fim temos a redução de dados, parte em que se descartam dados desnecessários para o estudo.(CAMILO; SILVA, 2009)

Após a preparação dos dados continua-se a mineração realizando algumas tarefas relevantes para o processo, como a descrição, que é a parte em que se descreve os padrões e tendências descobertos nos dados aplicando técnicas específicas de análise exploratória. Tem também a tarefa de classificação, que consiste em identificar a classe em que os dados pertencem, o que representam. Segue-se com a parte de estimação, sendo exclusiva para dados numéricos, onde se busca através do números gerar valores futuros, isso com a ajuda de algoritmos específicos, no caso de atributos não numéricos existe a predição que gera os valores desses dados. O agrupamento é outra tarefa a se seguir, nela agrupa-se os valores similares ou que tenha alguma complementação entre si, isso ajuda a enxergar melhor os dados. Finalizando as tarefas, tem-se a associação, que é nada mais que associar informações entre os dados, isso para construir e achar conhecimento.(CAMILO; SILVA, 2009)

Finalizando o processo de mineração é chegado o momento de se falar dos métodos, que são tecnicamente os algoritmos de mineração de dados, esses que são divididos em dois grupos,

os supervisionados e os não supervisionados. A diferença entre eles é que, os supervisionados são os métodos que utilizam conjuntos de dados de resultados já pré-definidos para trabalhar a saída de informações no treinamento. Já o não supervisionado os métodos tentam entender os dados sem nenhum resultado pré definido, trabalha de forma que com os dados o algoritmo chegue a algum resultado.(CAMILO; SILVA, 2009) Existem vários algoritmos que podem ser aplicados na mineração de dados, exemplos como, árvores de decisão, redes neurais, naive bayes, algoritmos genéticos, regras de associação, entre outros, vai da necessidade e da escolha do cientista de dados. Assim, através dos métodos é esperado chegar a conclusão do estudo e exploração dos dados, sendo que pode ou não ter encontrado conhecimento com a mineração.

4 OBJETIVOS

Objetivo Geral

O objetivo geral deste projeto é estudar os dados de pacientes de Covid-19 aplicando técnicas de mineração de dados com a intenção de encontrar características em comum entre os pacientes e relacionar os sintomas de Covid-19.

Objetivos Específicos

Os objetivos específicos do presente projeto são:

1. Obter os dados dos pacientes de Covid-19.

2. Fazer uma analise inicial dos dados obtidos.

3. Aplicar técnicas e algoritmos de mineração de dados para encontrar características e sintomas em comum entre os pacientes de Covid-19.

4. Descobrir conhecimentos nos dados de Covid-19.

5. Avaliar as informações descobertas.

6. Identificar problemas e mostrar resultados.

5 MATERIAIS E MÉTODOS

A base de dados utilizada para o estudo é uma base publica fornecida pelo Ministério da Saúde e disponibilizada no sistema SIVEP(Sistema de Informação de Vigilância Epidemiológica da Influenza), são dados de pacientes de covid-19 que deram entrada nos hospitais de todo o Brasil com sintomas de síndrome respiratória aguda grave durante o período de fevereiro de 2020 a abril de 2021. A base de dados trás informações como sexo, raça, sintomas, comorbidades, idade, dia de notificação do caso, escolaridade, localização, hospital, renda, fator de risco e entre outos, dos pacientes (MINISTERIODASAUDE, 2020-2021).

Foi escolhido para a Mineração de Dados o processo de KDD, que é definido como um processo onde-se realiza a descoberta de conhecimento em uma base de dados a partir da aplicação de técnicas específicas, essas que permitem que seja possível obter informações interessantes com os dados. A partir dos dados de pacientes de covid-19 foi usada essa técnica a fim de encontrar características e sintomas em comum entre os pacientes e assim podendo relacionar-los para a descoberta de algum conhecimento. Esse processo tem nove passos iterativos e interativos, ou seja, pode ser necessário voltar algumas etapas durante o processo para regular alguma etapa, por isso é necessário ter total conhecimento e clareza sobre todas as partes do processo. Segue-se nove passos na seguinte ordem: compreender o domínio dos dados, selecionar e criar o conjunto de dados, pré-processar e limpar os dados, transformar os dados, realizar a previsão e descrição, escolher o algoritmo de mineração de dados, utilizar o algoritmo de mineração, avaliar os resultados e, por fim, usar o conhecimento descoberto na base (MAIMON; ROKACH, 2005).

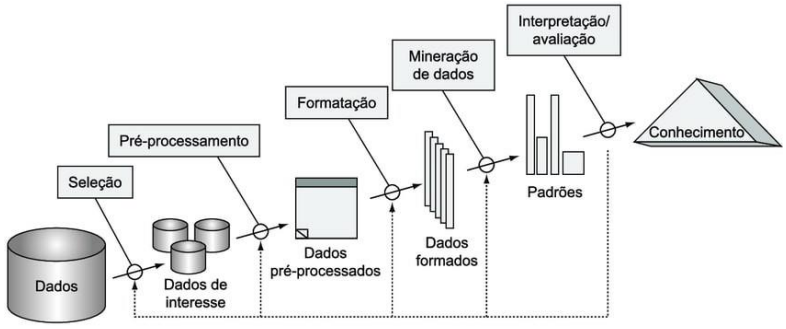


Figura 2 – Etapas do processo KDD (FAYYAD et al., 1996).

Como não se conhece a natureza da base de dados se inicia o processo criando um dicionário de dados para o entendimento inicial das informações contidas da base de pacientes de covid-19. O dicionário de dados é uma listagem dos elementos da base de dados de modo que fique bem definida a natureza deles. Com o dicionário de dados é possível obter várias informações sobre os dados, como tipo, tamanho, quantidade, domínio, descrição e o nome

dos atributos. O dicionário de dados auxilia na análise dos dados, servindo como um ponto inicial para realizar a seleção dos dados de interesse. Com a visualização dos dados das bases fica possível ver os elementos como os atributos e entidades, assim passa ser possível fazer a limpeza dos dados descartando os elementos desnecessários ou que estão em branco, deixando apenas os mais importantes para os estudos. Realiza-se então a formatação dos dados de modo que fica mais simples e prático trabalhar com o desejado.

Seguidamente com os dados formatados é realizada as estatísticas dos dados para que seja possível visualiza-los graficamente e assim ter uma ideai de como eles se comportam, para isso usa-se um algoritmo para gerar uma análise estatística dessa base. Pelo fato da base ser muito grande com alto volume de dados foi utilizado o Anaconda, que é um open source da linguagem de programação Python, para processar as informações e assim permitindo que seja possível gerar gráficos estatísticos dos dados. O Anaconda é uma ferramenta de distribuição python direcionada para área científica, especificamente para as áreas de ciência de dados, aprendizado de maquina, inteligência artificial e mineração de dados, ela permite a criação de algoritmos em python de forma direta e intuitiva. Tem uma grande distribuição de pacotes de bibliotecas específicas, aqui será escolhidas as direcionadas para manipulação dos dados e para a plotagem de gráficos estatísticos. Sendo uma delas a biblioteca Pandas que é uma ferramenta de análise e manipulação de dados de código aberto, rápida, poderosa, flexível e fácil de usar e a biblioteca matplotlib.pyplot fornece uma estrutura de plotagem de gráficos.

K-means foi uma das escolhas de algoritmo para a mineração de dados, ele é um algoritmo de clusterização onde se separa o conjunto de dados em vários grupos diferentes a partir de vários pontos (clusters), em que cada um desses pontos é de um subgrupo. Esse algoritmo possibilita ver e achar padrões nos dados através do agrupamento sendo uma escolha para o problema trabalhado, além de ser indicado para grandes conjunto de dados ele ainda vai possibilita maior visualização com pontos onde estão os problema (DABBURA, 2018). Porém no final o algoritmo utilizado foi o Apriori, pelo fato dele ser mais adequado para os dados trabalhados e para chegar nos resultados esperado. O A Priori é um algoritmo de classificação de conjuntos de itens frequentes que usa regras de associação para isso, recomendado para ser aplicado em grandes bases de dados com muitos números de transações entre os itens, ele é um ótimo algoritmo para se descobrir ou explorar padrões relevantes para estudos (AGRAWAL; IMIELIŃSKI; SWAMI, 1993).

6 ANÁLISE PRELIMINAR

Após os dados serem selecionados e pré-processados, foram gerados alguns gráficos e tabelas que permitiram a visualização da base, dessa forma foi possível analisar algumas informações contidas nos dados. Segue abaixo os gráficos e tabelas gerados:

- **Casos de Covid**

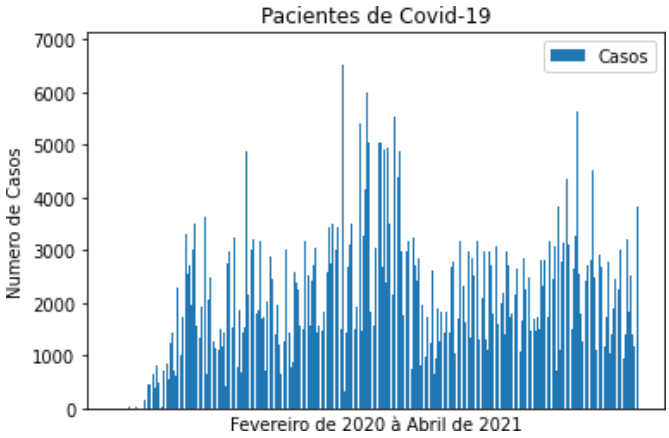


Figura 3 – **Gráfico de Barras dos caso de Covid-19**

Esse primeiro gráfico exibe o número de casos em relação ao decorrer dos dias, com o intervalo de tempo entre fevereiro de 2020 e abril de 2021. Observa-se os períodos de maior e de menor número de casos, assim sendo possível indicificar as ondas da pandemia (intervalos de crescimento acentuado de números de casos).

- **Perfil**

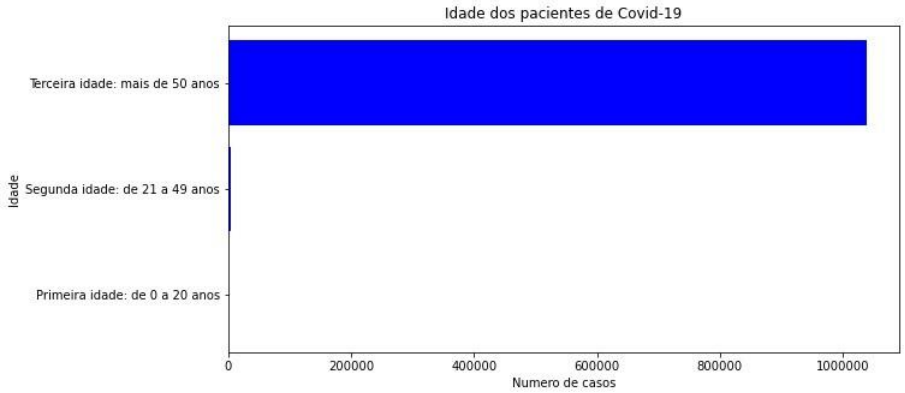


Figura 4 – **Gráfico de Barras verticais da idade dos pacientes.**

Agora analisando o perfil dos pacientes, vemos aqui na figura 4 o gráfico mostrando a idade dos pacientes, onde observamos que a grande maioria são da idade acima dos 50 anos, portanto confirma-se que idosos foram os mais afetados.

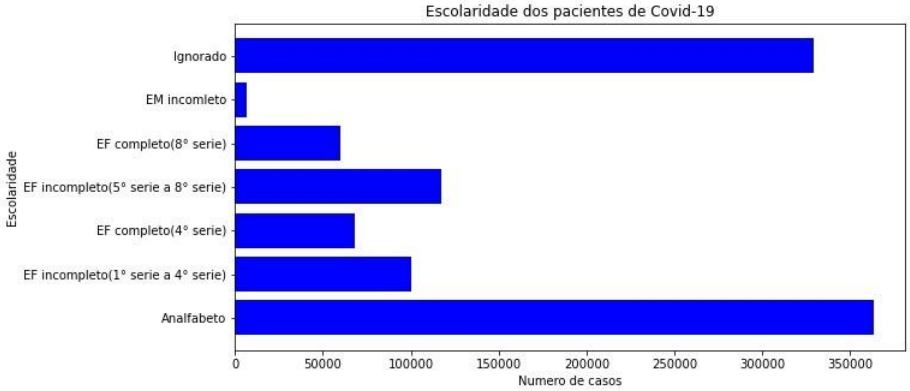


Figura 5 – Gráfico de Barras verticais da escolaridade dos pacientes.

Já aqui temos a escolaridade dos pacientes, indicando que grande parte é composta por pessoas analfabetas, e que muitos tiveram seus dados escolares ignorados na ficha de cadastro. Então relacionando com a figura 4, de idade, podemos concluir que a população idosa tem um índice muito alto de analfabetismo no Brasil.

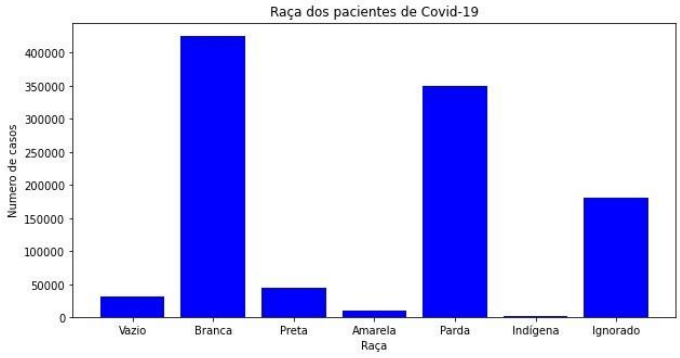


Figura 6 – Gráfico de Barras da raça dos pacientes.

Em relação as raças dos pacientes, vemos graficamente que a grande maioria dos pacientes são da raça branca seguidas da raça parda. Existe grande quantidade de pessoas que tiveram seus dados raciais ignorados. Assim concluímos que a raça branca é a que tem mais casos enquanto a preta e a indígena são as com menos.

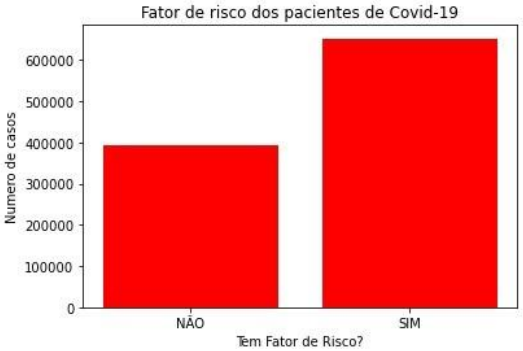


Figura 7 – Gráfico de Barras do fator de risco dos pacientes.

O gráfico trás a questão do fator de risco, onde concluímos que a maioria dos pacientes tem alguma característica que o coloca nesse campo de risco.

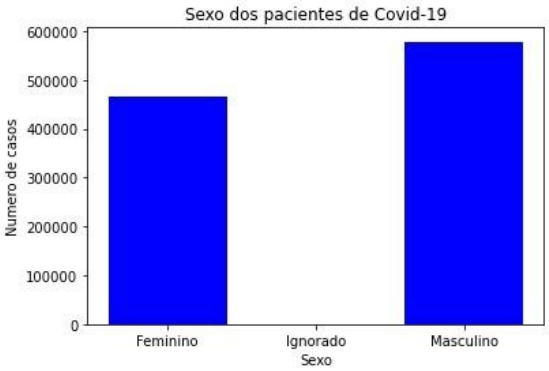


Figura 8 – Gráfico de Barras do sexo dos pacientes

Para finalizar o perfil, observamos o sexo dos pacientes, onde o gráfico mostra que entre os casos de covid-19 a maioria dos pacientes são do sexo masculino, afirmamos então que sexo feminino tem a menor taxa de pacientes.

- **Sintomas**

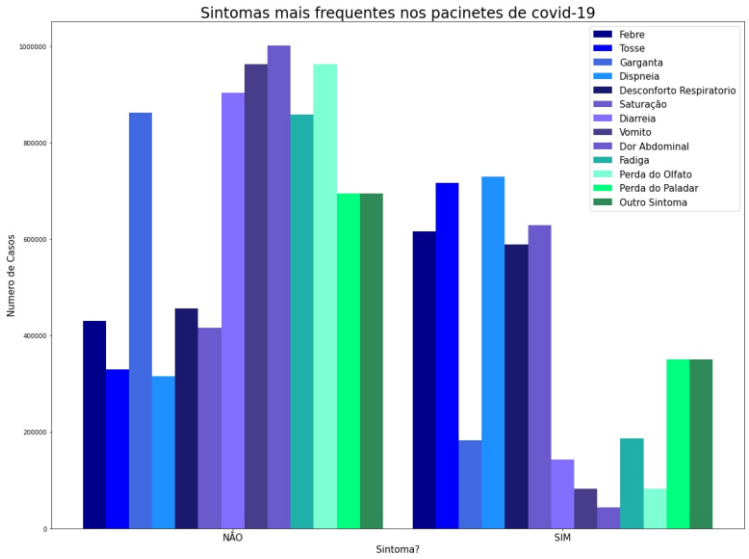


Figura 9 – Gráfico de Barras dos sintomas frequentes dos pacientes.

O gráfico acima mostra os sintomas mais frequentes entre os pacientes de covid-19, trazendo os casos em que os sintomas aparecem ou não. Observa-se que os sintomas mais frequentes são, em sequência, dispneia (falta de ar ou dificuldade de respirar), tosse, saturação (quantidade de oxigênio que está circulando no sangue) e febre. Ainda vemos que existem outros sintomas que são variados, esses que aparecem em quantidade muito baixa para serem frequentes e então são colocados em uma só categoria (Outro sintoma).

- **Comorbidades**

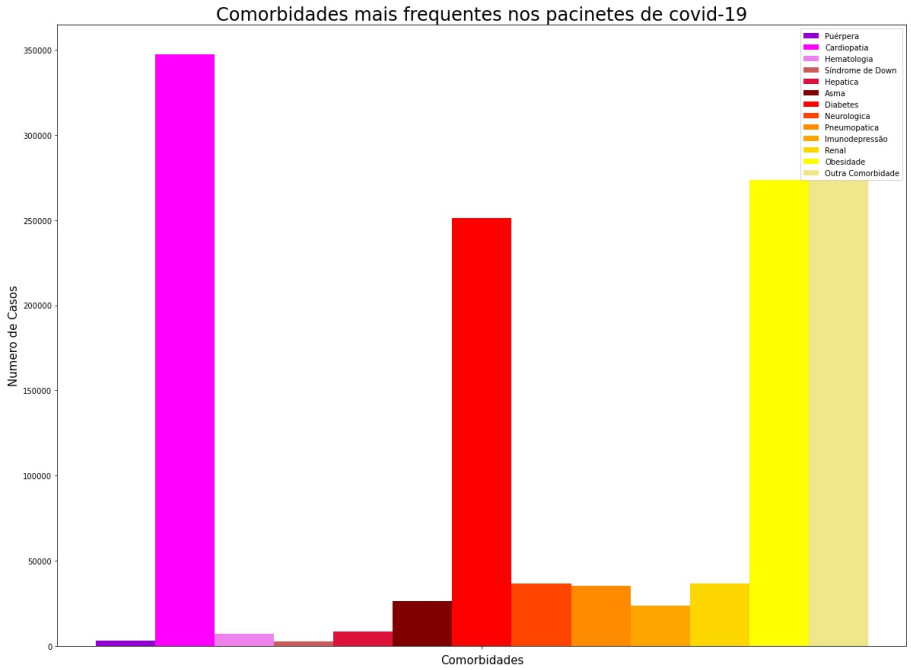


Figura 10 – **Gráfico de Barras das comorbidades frequentes.**

O gráfico trás as comorbidades que são mais comuns entre os casos de covid-19, onde podemos observar que cardiopatia é a comorbidade mais comum entre os pacientes, seguido de obesidade e diabetes. Temos também uma grande índice de variedades de outras comorbidades.

- **Correlações entre os itens**

Fazendo uma breve explicação do que é uma tabela de correlação, digamos que ela é um recurso de analise de dados para ver o quanto que dois itens de uma base de dados dependem um do outro para existir, ou seja, o quanto um interfere no outro para crescer ou decrescer. Esse valor é descoberto através do calculado do coeficiente Pearson utilizando a formula mostrada abaixo, essa formula calcula o valor de relação entre os itens, dando noção do grau de ligação.

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\left(\sum (x_i - \bar{x})^2\right)\left(\sum (y_i - \bar{y})^2\right)}}$$

Sabendo que de 1 á 0.7 o item está fortemente relacionado um com o outro, de 0.7 a 0.5 está moderadamente relacionado, de 0.5 a 0.2 tem pouca relação e abaixo disso aproximando de 0 tem nenhuma relação.

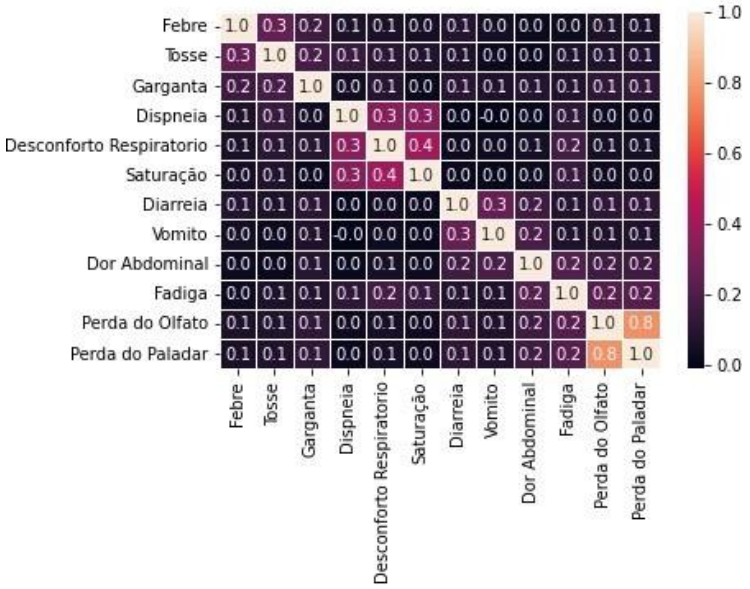


Tabela 1 – Tabela de correlação entre os itens de sintomas

A primeira tabela trás a correlação entre itens de sintomas, mostrando qual é o valor de conexão entre dois itens específicos da base.

Temos que perda de paladar e perda de olfato os itens mais relacionados entre si, com 0.8. Tendo também só que em grau baixo de relação tosse e febre, dispneia e desconforto respiratório, saturação e dispneia, desconforto respiratório e saturação, diarreia e vomito. Agora segunda tabela trás a correlação entre os sintomas.

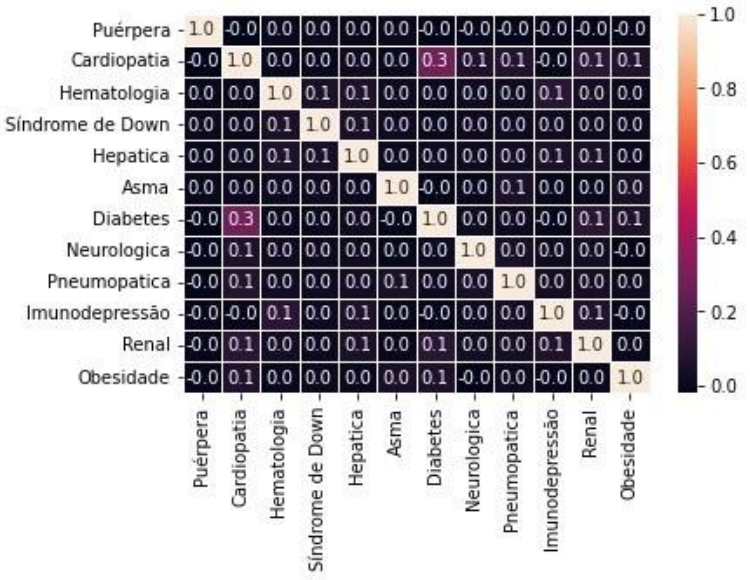


Tabela 2 – Tabela de correlação entre os itens de comorbidades

Observa-se que apenas que cardiopatia e diabetes tem correlação, com um grau baixo de 0.3.

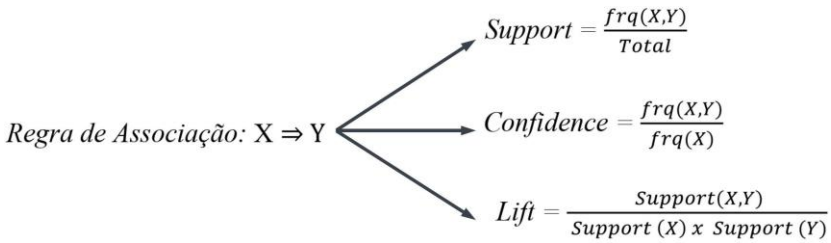
7 ALGORITMO APRIORI

O Algoritmo Apriori é um algoritmo de mineração de dados que trabalha com a ideia de conjuntos de itens frequentes, em outras palavras, ele ajuda a identificar os padrões que existem entre os dados. Desenvolvido em 1994 por R. Agrawal e R. Srikant, ele tem como objetivo encontrar conjuntos de itens frequentes em uma base de dados através de regras de associação. Regras de Associação tem como objetivo encontrar relacionamentos importantes entre grupos de itens nos dados trabalhados e exibe a frequência com que um conjunto de itens ocorre nesse processo de relacionamento (LIU, 2010).

- **Regra de Associação**

  Uma Regra de Associação é um método de descobrir a relação entre os itens de uma base de dados. Sabendo que uma base de dados é constituída de um conjunto de transações, temos que as transações são únicas e que cada uma é formada por um subconjunto de itens. Logo uma Regra de Associação é um relacionamento (X=>Y), sendo X e Y subconjuntos de itens, ou seja, se tem item X em um conjunto, provavelmente terá item Y também (AGRAWAL; IMIELIŃSKI; SWAMI, 1993).

  Para chegar em regras que sejam pertinentes para a base de dados são utilizados alguns cálculos específicos, sendo eles, Suporte(Support), Confiança(Confidence) e Aumento(Lift) (DONGRE; PRAJAPATI; TOKEKAR, 2014).



Support: é a proporção de um valor ou valores N na base de dados, ou seja, a frequência em que um determinado item ou itens aparecem no conjunto de dados.

Confidence: é o valor que informa a porcentagem de ocorrências em que a Regra de Associação é verdadeira, que é basicamente a frequência em que itens N aparecem juntos e associados em um conjunto.

Lift: é o valor que informa qual é o grau de associação entre os itens conjunto. Se o valor do lift for igual a 1 significa que os itens são independentes e não existe relação entre eles, se for acima de 1 os itens estão fortemente relacionados e se for menor que um os itens estão relacionados fracamente.

A escolha do nome Apriori para o algoritmo foi pelo fato dele usar o conhecimento prévio das propriedades frequentes do conjunto de itens, sabendo que a expressão A priori é utilizada no cotidiano com a finalidade de realizar menção a um motivo anterior à experiência. Na execução aplicamos uma abordagem iterativa ou busca em nível a onde conjuntos de itens k-frequentes são usados para encontrar conjuntos de itens k+1.

Com a intenção de aprimorar a eficiência da geração de conjuntos de itens frequentes, uma propriedade essencial é usada, que é nomeada de propriedade Apriori, que auxilia a reduzir o zona de busca (DONGRE; PRAJAPATI; TOKEKAR, 2014).

- **Propriedade Apriori**

  A propriedade Apriori afirma que, todo subconjunto não vazio do conjunto de itens frequente deve ser frequente. A definição base do algoritmo Apriori é sua anti-monotonicidade do valor de suporte. O algoritmo assume que todos os subconjuntos de um conjunto de itens frequente devem ser frequentes. Se um conjunto de itens for infrequente, todos os seus superconjuntos serão infrequentes.

Explicando agora o funcionamento e os passos do algoritmo será usado um conjunto de dados como exemplo que servirá como suporte para melhor entendimento. Esses dados se tratam de um lista de clientes de uma barraca de hortifrútis de uma feira, onde mostra os itens comprados por cada um desses clientes.

N° Cliente	Itens
1	Alface, Tomate, Cebola, Banana
2	Banana, Mamão, Abacaxi, Tomate
3	Tomate, Pimentão, Cebola, Alface
4	Tomate, Alface, Abacaxi, Cenoura
5	Tomate, Mamão, Alface, Banana
6	Alface, Pimentão, Cebola, Tomate, Abacaxi
7	Cebola, Banana, Pimentão, Repolho, Mamão
8	Alface, Pimentão, Cebola, Abacaxi
9	Alface, Repolho, Cebola, Pimentão, Tomate
10	Banana, Mamão, Tomate

Observando a base de dados vemos que ela é composta por duas colunas, onde a primeira é o numero do cliente e a segunda os produtos/itens comprados por cada um deles. Então o primeiro passo do Apriori vai ser calcular de forma individual o suporte de cada item da base, que vai ser basicamente contar o numero de ocorrências de um determinado produto e dividir pelo valor total, para isso utiliza-se a formula de suporte apresentada anteriormente, adaptando ao primeiro passo a formula será a seguinte, onde X é o item selecionado: $Support = \frac{Total}{frq(X)}$.

Assim geramos a tabela a seguir com o numero de suporte de cada produto.

Item	*Support*
Abacaxi	4/10 = 0,4
Alface	7/10 = 0,7
Banana	5/10 = 0,5
Cebola	6/10 = 0,6
Cenoura	1/10 = 0,1
Mamão	4/10 = 0,4
Pimentão	5/10 = 0,5
Repolho	2/10 = 0,2
Tomate	8/10 = 0,8

Depois de calcular o suporte de cada item X, passamos para o segundo passo, que é escolher o suporte mínimo para filtrar os itens que não são frequentes. Vamos escolher então para o exemplo trabalhado o peso 0,5 como critério de filtragem. Seguindo para o passo três, iremos usar o suporte mínimo e fazer a filtragem dos itens, logo temos os itens mais frequentes abaixo.

Item	*Support*
Alface	0,7
Banana	0,5
Cebola	0,6
Pimentão	0,5
Tomate	0,8

Indo para o quarto passo, será calculado novamente o suporte, mas agora para o conjunto de combinações de itens. Essa combinação é através das formações de pares de todos os itens do conjunto trabalhado, lembrando de ignorar as combinações onde aparecem itens não frequentes, esses que foram descartados pela filtragem. Após a combinação foi realizado o calculo de suporte novamente, agora adaptando a formula para dois itens (X e Y), a formulá ficara assim: $Support = \frac{frq(X,Y)}{Total}$.

Passando para o próximo passo, é feita a filtragem novamente com o suporte mínimo de 0,5 gerando assim um novo conjunto. Vemos então abaixo as duas tabelas geradas nessas duas etapas, sendo a primeira o resultado das combinações e o calculo de suporte dos itens e a segunda a filtragem com o suporte mínimo.

Itens combinados	*Support*
Alface, Banana	2/10 = 0,2
Alface, Cebola	5/10 = 0,5
Alface, Pimentão	4/10 = 0,4
Alface, Tomate	6/10 = 0,6
Banana, Cebola	2/10 = 0,2
Banana, Pimentão	1/10 = 0,1
Banana, Tomate	4/10 = 0,4
Cebola, Pimentão	5/10 = 0,5
Cebola, Tomate	4/10 = 0,4
Pimentão, Tomate	3/10 = 0,3

Itens combinados	*Support*
Alface, Cebola	0,5
Alface, Tomate	0,6
Cebola, Pimentão	0,5

No sexto passo é repetido processo para combinações, só que agora entre três itens, lembrando de descartar as duplas não frequentes para formar os trios chegamos as combinações e já é feito novamente o calculo de suporte, adaptando a formula para três tipos de itens, sendo a seguinte formula: $Support = \frac{frq(X,Y,Z)}{Total}$. Chegamos aos seguintes conjuntos de trios abaixo com seus suportes calculados.

Itens combinados	*Support*
Alface, Cebola, Pimentão	4/10 = 0,4
Alface, Tomate, Cebola	4/10 = 0,4

Filtrando os trios de conjuntos onde o desejado é que suporte mínimo seja de 0,5, descartamos todos os dois trios, já que eles tem o suporte 0,4. Então prosseguimos os passos do algoritmo com os conjuntos de duplas apresentados anteriormente, pous apresentam o suporte mínimo desejado.

Itens combinados	*Support*
Alface, Cebola	0,5
Alface, Tomate	0,6
Cebola, Pimentão	0,5

O próximo e ultimo passo do algoritmo Apriore é criar as Regras de Associação e calcular a confiança e o lift dos conjuntos.

Itens combinados	*Support*	*Confidence*	*Lift*
Alface, Cebola	0,5	5/7=0,71	0,5/(0,7x0,6) =1,19
Alface, Tomate	0,6	6/7=0,85	0,6/(0,7x0,8) = 1,07
Cebola, Pimentão	0,5	5/6=0,83	0,5/(0,6x0,5) = 1,66

Finalizando o algoritmo temos o seguinte resultado:

Itens combinados	*Regra*	*Support*	*Confidence*	*Lift*
Alface, Cebola	Alface ⇒ Cebola	0,5	0,71	1,19
Alface, Tomate	Alface ⇒Tomate	0,6	0,85	1,07
Cebola, Pimentão	Cebola ⇒ Pimentão	0,5	0,83	1,66

Com o algoritmo Apriori chegamos á alguns resultados a cerca da base de dados da barraca de hortifrúti da feira, sendo eles que:

- 50% dos clientes que compraram alface levaram cebola;

  A confiança de uma pessoa que compre alface levar cebola é de 71%.

- 60% dos clientes que compraram alface levaram tomate;

  A confiança de uma pessoa que compre alface levar tomate é de 85%.

- 50% dos clientes que compraram cebola levaram pimentão;

  A confiança de uma pessoa que compre cebola levar pimentão é de 83%.

8 RESULTADOS

Escolher o algoritmo de mineração de dados foi uma tarefa desafiadora, já que existe uma variedade muito grande de algoritmos e poucos apenas que podem funcionar nos dados trabalhados, por isso nesse momento é preciso saber escolher o que mais vai gerar resultados consistentes e uteis para o estudo. Após a formatação e uma analise profunda dos dados e dos gráficos gerados concluiu-se que o Apriori seria o mais adequado para a base trabalhada pois ele funciona com a ideia de itens frequentes e como a intenção principal aqui é relacionar as caraterísticas dos pacientes deu muito certo.

A implementação do algoritmo foi feita em Python utilizando a plataforma Anaconda que disponibiliza varias bibliotecas importantes e uteis para trabalhar com data mining. Instalando o pacote do Apriori é só adequar os dados para serem treinados no algoritmo, assim após fazer a separação dos dados foram treinamos cada grupo separados da base de dados e chegamos a alguns resultados interessantes que serão mostrados a seguir.

- **Sintomas**

  O primeiro resultado trás o algoritmo aplicado aos dados de sintomas dos pacientes de covid-19, após a separação dos dados e uma analise dos gráficos de frequência foi escolhido o valor de 0.45 para ser o suporte mínimo para a filtragem do algoritmo, temos como resultado a figura 11.

antecedents	consequents	antecedent support	consequent support	support	confidence	lift
(Desconforto Respiratorio)	(Dispneia)	0.563234	0.698255	0.470789	0.835868	1.197081
(Saturação)	(Dispneia)	0.601374	0.698255	0.496497	0.825604	1.182383
(Febre)	(Tosse)	0.588768	0.685039	0.462655	0.785802	1.147090
(Tosse)	(Dispneia)	0.685039	0.698255	0.509036	0.743075	1.064189
(Dispneia)	(Tosse)	0.698255	0.685039	0.509036	0.729011	1.064189

Figura 11 – **Resultado do algoritmo Apriori treinado com os dados de sintomas dos pacientes de covid-19**

Analisando o resultado vemos que encontramos 5 regras de associação para os sintomas, sendo elas: **Desconforto respiratório** => **Dispneia** com o support igual a 0.470789, confidence 0.835868 e o lift de 1.197081; **Saturação** => **Dispneia** com o support igual a 0.496497, confidence 0.825604 e o lift de 1.182383; **Febre** => **Tosse** com o support igual a 0.462655, confidence 0.785802 e o lift de 1.147090; **Tosse** => **Dispneia** com o support igual a 0.509036, confidence 0.743075 e o lift de 1.064189; **Dispneia** => **Tosse** com o support igual a 0.509036, confidence 0.729011 e o lift de 1.064189.

Separando os pacientes por sexo (Feminino e Masculino) descartando os que não informaram o sexo, temos os dois resultados mostrados respectivamente nas figuras 12 e

13 onde o suporte continua com mínimo de 0.45 pois esse valor está sendo adequado para alcançar resultados desejados com esse grupo de dados separado da base de dados.

antecedents	consequents	antecedent support	consequent support	support	confidence	lift
(Desconforto Respiratorio)	(Dispneia)	0.557978	0.692299	0.465483	0.834231	1.205015
(Saturação)	(Dispneia)	0.591434	0.692299	0.487560	0.824370	1.190771
(Tosse)	(Dispneia)	0.672936	0.692299	0.498138	0.740245	1.069255
(Dispneia)	(Tosse)	0.692299	0.672936	0.498138	0.719541	1.069255
(Dispneia)	(Saturação)	0.692299	0.591434	0.487560	0.704262	1.190771

Figura 12 – **Resultado do algoritmo Apriori treinado com os dados de sintomas de covid-19 dos pacientes do sexo feminino**

Analisando o resultado dos pacientes do sexo feminino encontramos 5 regras de associação para os sintomas, sendo elas: **Desconforto respiratório => Dispneia** com o support igual a 0.465483, confidence 0.834231 e o lift de 1.205015; **Saturação => Dispneia** com o support igual a 0.487560, confidence 0.824370 e o lift de 1.190771; **Tosse => Dispneia** com o support igual a 0.498138, confidence 0.740245 e o lift de 1.069255; **Dispneia => Tosse** com o support igual a 0.498138, confidence 0.719541 e o lift de 1.069255; **Dispneia => Saturação** com o support igual a 0.487560, confidence 0.704262 e o lift de 1.190771.

antecedents	consequents	antecedent support	consequent support	support	confidence	lift
(Desconforto Respiratorio)	(Dispneia)	0.567434	0.703045	0.475030	0.837155	1.190755
(Saturação)	(Dispneia)	0.609355	0.703045	0.503665	0.826555	1.175679
(Febre)	(Tosse)	0.621532	0.694805	0.490067	0.788483	1.134826
(Tosse)	(Dispneia)	0.694805	0.703045	0.517822	0.745277	1.060070
(Dispneia)	(Tosse)	0.703045	0.694805	0.517822	0.736542	1.060070

Figura 13 – **Resultado do algoritmo Apriori treinado com os dados de sintomas de covid-19 dos pacientes do sexo masculino**

Agora o resultado dos pacientes do sexo masculino onde encontramos 5 regras de associação para os sintomas, sendo elas: **Desconforto respiratório => Dispneia** com o support igual a 0.475030, confidence 0.837155 e o lift de 1.190755; **Saturação => Dispneia** com o support igual a 0.503665, confidence 0.826555 e o lift de 1.175679; **Febre => Tosse** com o support igual a 0.490067, confidence 0.788483 e o lift de 1.134826; **Tosse => Dispneia** com o support igual a 0.517822, confidence 0.745277 e o lift de 1.060070; **Dispneia => Tosse** com o support igual a 0.517822, confidence 0.736542 e o lift de 1.060070.

Separando aqui os pacientes por raça, sendo que após uma analise dos dados e dos gráficos foi decidido apenas selecionar os pacientes das raças Preta, Parda e Branca pois são as três mais frequentes no conjunto de dados, ainda descartando os pacientes que

não teve suas raças declaradas no registro do caso de covid-19, ainda utilizando o suporte mínimo de 0.45 temos têS resultados apresentados nas figuras 14, 15 e 16.

antecedents	consequents	antecedent support	consequent support	support	confidence	lift
(Desconforto Respiratorio)	(Dispneia)	0.590360	0.708050	0.494266	0.837228	1.182443
(Saturação)	(Dispneia)	0.625752	0.708050	0.518652	0.828845	1.170603
(Febre)	(Tosse)	0.573641	0.690848	0.457753	0.797978	1.155071
(Desconforto Respiratorio)	(Saturação)	0.590360	0.625752	0.462806	0.783939	1.252794
(Tosse)	(Dispneia)	0.690848	0.708050	0.524122	0.758666	1.071487

Figura 14 – **Resultado do algoritmo Apriori treinado com os dados de sintomas de covid-19 dos pacientes da raça Preta**

Observando os pacientes da raça preta encontramos 5 regras de associação para os sintomas, sendo elas: **Desconforto respiratório => Dispneia** com o support igual a 0.494266, confidence 0.837228 e o lift de 1.182443; **Saturação => Dispneia** com o support igual a 0.518652, confidence 0.828845 e o lift de 1.170603; **Febre => Tosse** com o support igual a 0.457753, confidence 0.797978 e o lift de 1.155071; **Desconforto respiratório => Saturação** com o support igual a 0.462806, confidence 0.783939 e o lift de 1.252794; **Tosse => Dispneia** com o support igual a 0.524122, confidence 0.758666 e o lift de 1.071487.

antecedents	consequents	antecedent support	consequent support	support	confidence	lift
(Desconforto Respiratorio)	(Dispneia)	0.567434	0.703045	0.475030	0.837155	1.190755
(Saturação)	(Dispneia)	0.609355	0.703045	0.503665	0.826555	1.175679
(Febre)	(Tosse)	0.621532	0.694805	0.490067	0.788483	1.134826
(Tosse)	(Dispneia)	0.694805	0.703045	0.517822	0.745277	1.060070
(Dispneia)	(Tosse)	0.703045	0.694805	0.517822	0.736542	1.060070

Figura 15 – **Resultado do algoritmo Apriori treinado com os dados de sintomas de covid-19 dos pacientes da raça Parda**

Agora o resultado dos pacientes da raça Parda onde encontramos 5 regras de associação para os sintomas, sendo elas: **Saturação => Dispneia** com o support igual a 0.496941, confidence 0.844738 e o lift de 1.183988; **Desconforto Respiratório => Dispneia** com o support igual a 0.499082, confidence 0.842438 e o lift de 1.180764; **Febre => Tosse** com o support igual a 0.494346, confidence 0.806533 e o lift de 1.145841; **Saturação => Desconforto Respiratório** com o support igual a 0.451869, confidence 0.768121 e o lift de 1.296568; **Desconforto Respiratório => Saturação** com o support igual a 0.451869, confidence 0.762743 e o lift de 1.296568.

Nos dados de pacientes da raça Branca encontramos também 5 regras de associação para os sintomas, sendo elas: **Desconforto respiratório => Dispneia** com o support igual a 0.493476, confidence 0.840135 e o lift de 1.184520; **Saturação => Dispneia** com o support igual a 0.531090, confidence 0.824482 e o lift de 1.162451; **Desconforto respiratório => Saturação** com o support igual a 0.466238, confidence 0.793763 e o lift de

antecedents	consequents	antecedent support	consequent support	support	confidence	lift
(Desconforto Respiratorio)	(Dispneia)	0.587377	0.709262	0.493476	0.840135	1.184520
(Saturação)	(Dispneia)	0.644149	0.709262	0.531090	0.824482	1.162451
(Desconforto Respiratorio)	(Saturação)	0.587377	0.644149	0.466238	0.793763	1.232266
(Dispneia)	(Saturação)	0.709262	0.644149	0.531090	0.748791	1.162451
(Tosse)	(Dispneia)	0.684730	0.709262	0.511001	0.746281	1.052193

Figura 16 – **Resultado do algoritmo Apriori treinado com os dados de sintomas de covid-19 dos pacientes da raça Branca**

1.232266; **Dispneia => Saturação** com o support igual a 0.531090, confidence 0.748791 e o lift de 1.162451; **Tosse => Dispneia** com o support igual a 0.511001, confidence 0.746281 e o lift de 1.052193.

Indo mais a fundo nos dados foi feita a seleção dos grupos levando em consideração o sexo e a raça, assim como anteriormente apenas as raças Preta, Pardas e Branca foram selecionadas para o treinamento do Apriori, pacientes do sexo feminino das raças preta, parda e branca, e pacientes do sexo masculino das raças preta, parda e branca, os resultados estão nas figuras 17, 18, 19, 20, 21 e 22, ainda com o suporte mínimo de 0.45.

antecedents	consequents	antecedent support	consequent support	support	confidence	lift
(Desconforto Respiratorio)	(Dispneia)	0.586127	0.710667	0.493839	0.842546	1.185571
(Saturação)	(Dispneia)	0.624024	0.710667	0.519660	0.832756	1.171795
(Desconforto Respiratorio)	(Saturação)	0.586127	0.624024	0.458495	0.782245	1.253548
(Tosse)	(Dispneia)	0.676943	0.710667	0.518040	0.765265	1.076826
(Saturação)	(Desconforto Respiratorio)	0.624024	0.586127	0.458495	0.734739	1.253548

Figura 17 – **Resultado do algoritmo Apriori treinado com os dados de sintomas de covid-19 dos pacientes do sexo feminino e da raça preta**

Observa-se no resultado do grupo de pacientes do sexo feminino e raça preta 5 regras de associação para os sintomas, sendo elas: **Desconforto respiratório => Dispneia** com o support igual a 0.493839, confidence 0.842546 e o lift de 1.185571; **Saturação => Dispneia** com o support igual a 0.519660, confidence 0.832756 e o lift de 1.171795; **Desconforto respiratório => Saturação** com o support igual a 0.458495, confidence 0.782245 e o lift de 1.253548; **Tosse => Dispneia** com o support igual a 0.518040, confidence 0.765265 e o lift de 1.076826; **Saturação => Desconforto Respiratório** com o support igual a 0.458495, confidence 0.734739 e o lift de 1.253548.

O resultado do grupo dos pacientes do sexo feminino e raça parda trás 5 regras de associação para os sintomas, elas são: **Saturação => Dispneia** com o support igual a 0.485893, confidence 0.843387 e o lift de 1.194963; **Desconforto respiratório => Dispneia** com o support igual a 0.491113, confidence 0.839683 e o lift de 1.189715 ; **Febre => Tosse** com o support igual a 0.465176, confidence 0.803692 e o lift de 1.159084; **Tosse => Dispneia** com o support igual a 0.522108, confidence 0.752983 e o lift de 1.066873;

antecedents	consequents	antecedent support	consequent support	support	confidence	lift
(Saturação)	(Dispneia)	0.576122	0.705785	0.485893	0.843387	1.194963
(Desconforto Respiratorio)	(Dispneia)	0.584879	0.705785	0.491113	0.839683	1.189715
(Febre)	(Tosse)	0.578799	0.693386	0.465176	0.803692	1.159084
(Tosse)	(Dispneia)	0.693386	0.705785	0.522108	0.752983	1.066873
(Dispneia)	(Tosse)	0.705785	0.693386	0.522108	0.739754	1.066873

Figura 18 – **Resultado do algoritmo Apriori treinado com os dados de sintomas de covid-19 dos pacientes do sexo feminino e da raça parda**

Dispneia => Tosse com o support igual a 0.522108, confidence 0.739754 e o lift de 1.066873.

antecedents	consequents	antecedent support	consequent support	support	confidence	lift
(Desconforto Respiratorio)	(Dispneia)	0.582642	0.704802	0.488581	0.838561	1.189783
(Saturação)	(Dispneia)	0.636274	0.704802	0.523211	0.822305	1.166718
(Desconforto Respiratorio)	(Saturação)	0.582642	0.636274	0.459171	0.788084	1.238594
(Tosse)	(Dispneia)	0.672308	0.704802	0.500715	0.744771	1.056710
(Dispneia)	(Saturação)	0.704802	0.636274	0.523211	0.742352	1.166718

Figura 19 – **Resultado do algoritmo Apriori treinado com os dados de sintomas de covid-19 dos pacientes do sexo feminino e da raça branca**

Passando para o resultado do grupo de pacientes do sexo feminino e raça branca entramos 5 regras de associação para os sintomas, elas são: **Desconforto respiratório => Dispneia** com o support igual a 0.488581, confidence 0.838561 e o lift de 1.189783; **Saturação => Dispneia** com o support igual a 0.523211, confidence 0.822305 e o lift de 1.166718; **Desconforto respiratório => Saturação** com o support igual a 0.459171, confidence 0.788084 e o lift de 1.238594; **Tosse => Dispneia** com o support igual a 0.500715, confidence 0.744771 e o lift de 1.056710; **Dispneia => Tosse** com o support igual a 0.523211, confidence 0.742352 e o lift de 1.166718.

antecedents	consequents	antecedent support	consequent support	support	confidence	lift
(Desconforto Respiratorio)	(Dispneia)	0.586127	0.710667	0.493839	0.842546	1.185571
(Saturação)	(Dispneia)	0.624024	0.710667	0.519660	0.832756	1.171795
(Desconforto Respiratorio)	(Saturação)	0.586127	0.624024	0.458495	0.782245	1.253548
(Tosse)	(Dispneia)	0.676943	0.710667	0.518040	0.765265	1.076826
(Saturação)	(Desconforto Respiratorio)	0.624024	0.586127	0.458495	0.734739	1.253548

Figura 20 – **Resultado do algoritmo Apriori treinado com os dados de sintomas de covid-19 dos pacientes do sexo masculino e da raça preta**

Observa-se na figura 20 o resultado do grupo de pacientes do sexo masculino e raça preta 5 regras de associação para os sintomas, sendo elas: **Desconforto respiratório => Dispneia** com o support igual a 0.494534, confidence 0.832943 e o lift de 1.179901; **Saturação => Dispneia** com o support igual a 0.517829, confidence 0.825632 e o lift de 1.169545; **Febre => Tosse** com o support igual a 0.484914, confidence 0.803187 e o lift de 1.143999; **Desconforto Respiratório => Saturação** com o support igual a 0.466269,

confidence 0.785336 e o lift de 1.252148; **Tosse => Dispneia** com o support igual a 0.529040, confidence 0.753525 e o lift de 1.067401.

antecedents	consequents	antecedent support	consequent support	support	confidence	lift
(Saturação)	(Dispneia)	0.576122	0.705785	0.485893	0.843387	1.194963
(Desconforto Respiratorio)	(Dispneia)	0.584879	0.705785	0.491113	0.839683	1.189715
(Febre)	(Tosse)	0.578799	0.693386	0.465176	0.803692	1.159084
(Tosse)	(Dispneia)	0.693386	0.705785	0.522108	0.752983	1.066873
(Dispneia)	(Tosse)	0.705785	0.693386	0.522108	0.739754	1.066873

Figura 21 – **Resultado do algoritmo Apriori treinado com os dados de sintomas de covid-19 dos pacientes do sexo masculino e da raça parda**

O resultado do grupo dos pacientes do sexo masculino e raça parda trás 5 regras de associação para os sintomas, elas são: **Saturação => Dispneia** com o support igual a 0.505627, confidence 0.845745 e o lift de 1.175427; **Desconforto respiratório => Dispneia** com o support igual a 0.505325, confidence 0.844559 e o lift de 1.173779; **Febre => Tosse** com o support igual a 0.517435, confidence 0.808558 e o lift de 1.135321; **Saturação => Desconforto Respiratório** com o support igual a 0.460195, confidence 0.769752 e o lift de 1.286501; **Desconforto Respiratório => Saturação** com o support igual a 0.460195, confidence 0.769133 e o lift de 1.286501.

antecedents	consequents	antecedent support	consequent support	support	confidence	lift
(Desconforto Respiratorio)	(Dispneia)	0.582642	0.704802	0.488581	0.838561	1.189783
(Saturação)	(Dispneia)	0.636274	0.704802	0.523211	0.822305	1.166718
(Desconforto Respiratorio)	(Saturação)	0.582642	0.636274	0.459171	0.788084	1.238594
(Tosse)	(Dispneia)	0.672308	0.704802	0.500715	0.744771	1.056710
(Dispneia)	(Saturação)	0.704802	0.636274	0.523211	0.742352	1.166718

Figura 22 – **Resultado do algoritmo Apriori treinado com os dados de sintomas de covid-19 dos pacientes do sexo masculino e da raça branca**

Enfim por ultimo temos o resultado do grupo de pacientes do sexo masculino e raça branca entramos 5 regras de associação para os sintomas, elas são: **Desconforto respiratório => Dispneia** com o support igual a 0.497536, confidence 0.841424 e o lift de 1.180165; **Saturação => Dispneia** com o support igual a 0.537620, confidence 0.826241 e o lift de 1.158870; **Desconforto respiratório => Saturação** com o support igual a 0.472101, confidence 0.798409 e o lift de 1.227037; **Febre => Tosse** com o support igual a 0.473676, confidence 0.781118 e o lift de 1.123790; **Dispneia => Saturação** com o support igual a 0.537620, confidence 0.754055 e o lift de 1.158870;

- **Comorbidades**

  Vindo agora para os dados de comorbidades chegamos a uma questão, que é a quantidade de dados, se tem pouco sobre as comorbidades, sendo que sua presença é bem baixa nos casos, assim dificultando alguma analise mais profunda dos dados. Porem usando um valor bem baixo de suporte o algoritmo chegou a um resultado que será mostrado na figura 23, o suporte mínimo foi de 0.10.

antecedents	consequents	antecedent support	consequent support	support	confidence	lift
(Diabetes)	(Cardiopatia)	0.240484	0.332313	0.13671	0.568479	1.710674
(Cardiopatia)	(Diabetes)	0.332313	0.240484	0.13671	0.411389	1.710674

Figura 23 – **Resultado do algoritmo Apriori treinado com os dados de comorbidade dos pacientes de covid-19**

Temos então duas regras, que são: **Diabetes => Cardiopatia** com o support igual a 0.13671, confidence 0.568479 e o lift de 1.710674; **Cardiopatia => Diabetes** com o support igual a 0.13671, confidence 0.411389 e o lift de 1.710674;

9 CONCLUSÕES

Com os resultados foi possível chegar em algumas conclusões bastante interessantes acerca dos dados da base, principalmente a relação entre os sintomas e as características dos pacientes de covid-19, que foi o foco deste projeto.

Através dos gráficos estatísticos dos dados chegamos em alguns pontos importantes, que são, a maioria dos pacientes são pessoas idosas; que os incidente de casos foram maiores em pessoas do sexo masculino e de raça branca; que dispneia, tosse, saturação e febre foram os sintomas que mais ocorreram; e por ultimo que pessoas com comorbidades do tipo obesidade e diabética tiveram maior vulnerabilidade.

Para finalizar temos os gráficos que mostra os conhecimentos descobertos. Primeiro temos as relações entre os sintomas descobertas através do Apriori.

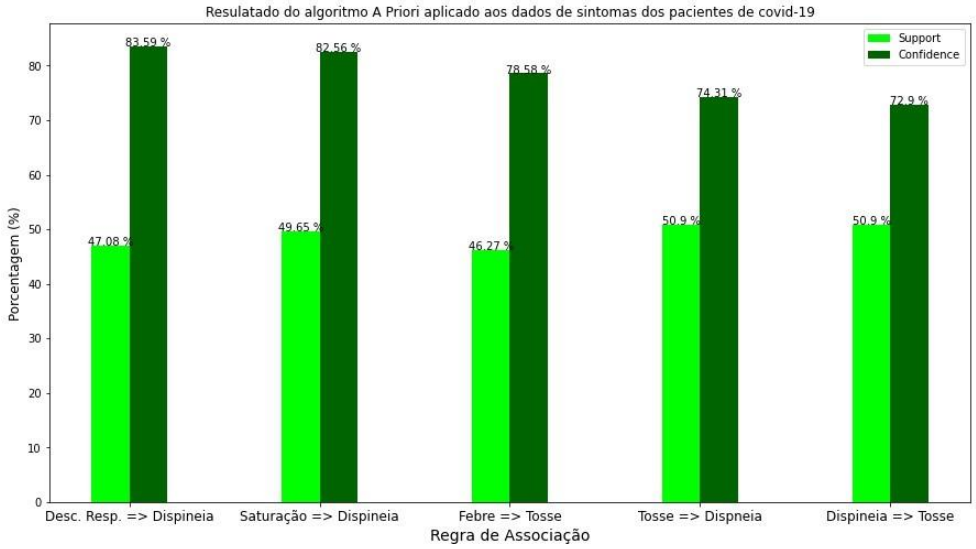


Figura 24 – Gráfico com o resultado do Apriori aplicado aos sintomas de covid-19

- 47,08% dos pacientes que tiveram desconforto respiratório manifestaram também dispneia, sendo a confiança dessa regra acorrer de 83,59%;
- 49,65% dos pacientes que tiveram saturação manifestaram também dispneia, sendo a confiança dessa regra acorrer de 82,56%;
- 46,27% dos pacientes que tiveram febre manifestaram também tosse, sendo a confiança dessa regra acorrer de 78,58%;
- 50,9% dos pacientes que tiveram tosse manifestaram também dispneia, sendo a confiança dessa regra acorrer de 74,31%;

- 50,9% dos pacientes que tiveram dispneia manifestaram também tosse, sendo a confiança dessa regra acorrer de 72,9%;

O maior índice de ocorrência de dois sintomas relacionados é de que pessoas que tiveram tosse também apresentaram dispneia, mas com a confiança podemos afirmar que a associação mais garantida entre os sintomas é que o paciente que tiver desconforto respiratório certamente terá dispneia.

Agora relacionando o sexo dos pacientes com os sintomas.

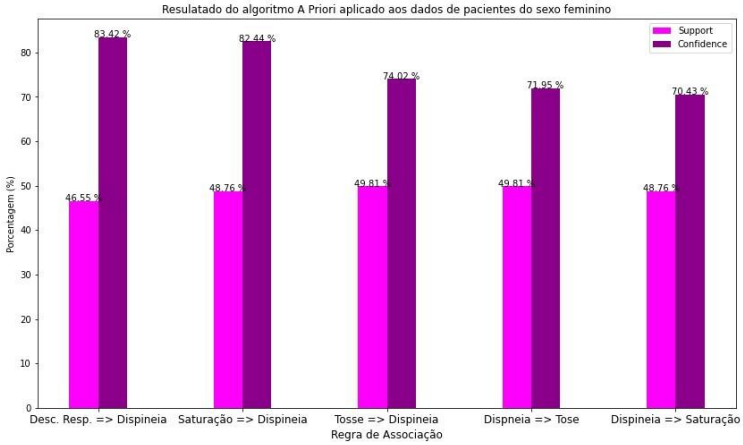


Figura 25 – **Resultado gráfico do Apriori relacionando os sintomas ao sexo feminino**

- 46,55% dos pacientes do sexo feminino que tiveram desconforto respiratório manifestaram também dispneia, sendo a confiança dessa regra acorrer de 83,42%;

- 48,76% dos pacientes do sexo feminino que tiveram saturação manifestaram também dispneia, sendo a confiança dessa regra acorrer de 82,44%;

- 49,81% dos pacientes do sexo feminino que tiveram tosse manifestaram também dispneia, sendo a confiança dessa regra acorrer de 74,02%;

- 49,81% dos pacientes do sexo feminino que tiveram dispneia manifestaram também tosse, sendo a confiança dessa regra acorrer de 71,95%;

- 48,76% dos pacientes do sexo feminino que tiveram dispneia manifestaram também saturação, sendo a confiança dessa regra acorrer de 70,43%;

O maior índice de ocorrência de dois sintomas relacionados entre os pacientes do sexo feminino é que quem tiver tosse também apresentaram dispneia, mas com a confiança podemos afirmar que a associação mais provável de ocorrer entre os sintomas é que o paciente que tiver desconforto respiratório certamente terá dispneia.

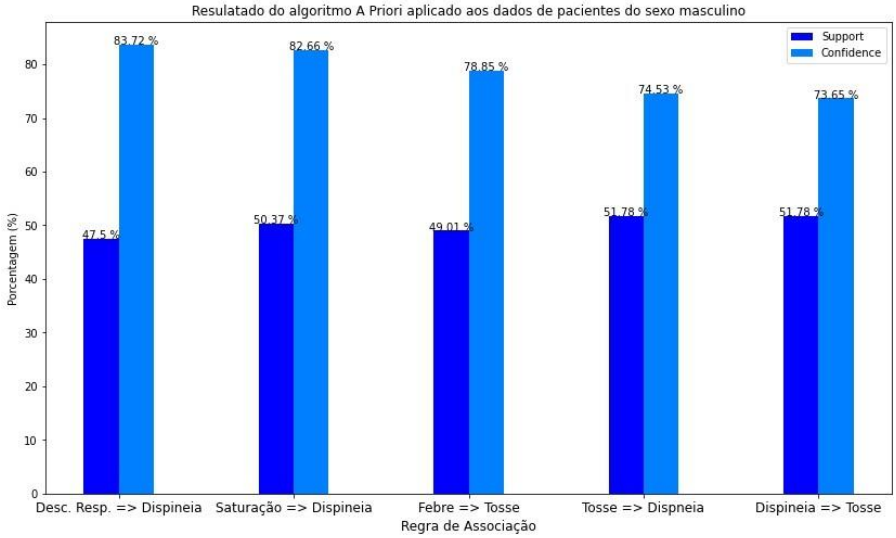


Figura 26 – **Resultado gráfico do Apriori relacionando os sintomas ao sexo masculino**

- 47,5% dos pacientes do sexo masculino que tiveram desconforto respiratório manifestaram também dispneia, sendo a confiança dessa regra acorrer de 83,72%;
- 50,37% dos pacientes do sexo masculino que tiveram saturação manifestaram também dispneia, sendo a confiança dessa regra acorrer de 82,66%;
- 49,01% dos pacientes do sexo masculino que tiveram febre manifestaram tosse, sendo a confiança dessa regra acorrer de 78,85%;
- 51,78% dos pacientes do sexo masculino que tiveram tosse manifestaram dispneia, sendo a confiança dessa regra acorrer de 74,53%;
- 51,78% dos pacientes do sexo masculino que tiveram dispneia manifestaram também tosse, sendo a confiança dessa regra acorrer de 73,65%;

Entre os pacientes do sexo masculino a associação que teve maior índice de ocorrência foi a quem teve tosse também manifestou dispneia e valendo também para a associação ao contrario (dispneia=>tosse), porém quando olhamos para a confiança podemos afirmar que a associação mais garantida de ocorrer entre os relacionamentos é a de que se o paciente do sexo masculino tiver desconforto respiratório certamente terá dispneia também, assim como nos pacientes de sexo feminino.

Logo podemos dizer acerca dos dados apresentados que o sexo tem pouca influencia nas relações entre sintomas, só se pode afirmar que pessoas do sexo masculino tem maior manifestação de sintomas do que as pessoas de sexo feminino.

Vamos agora mostrar os resultados do Apriori quando relacionamos os sintomas com a raça dos pacientes.

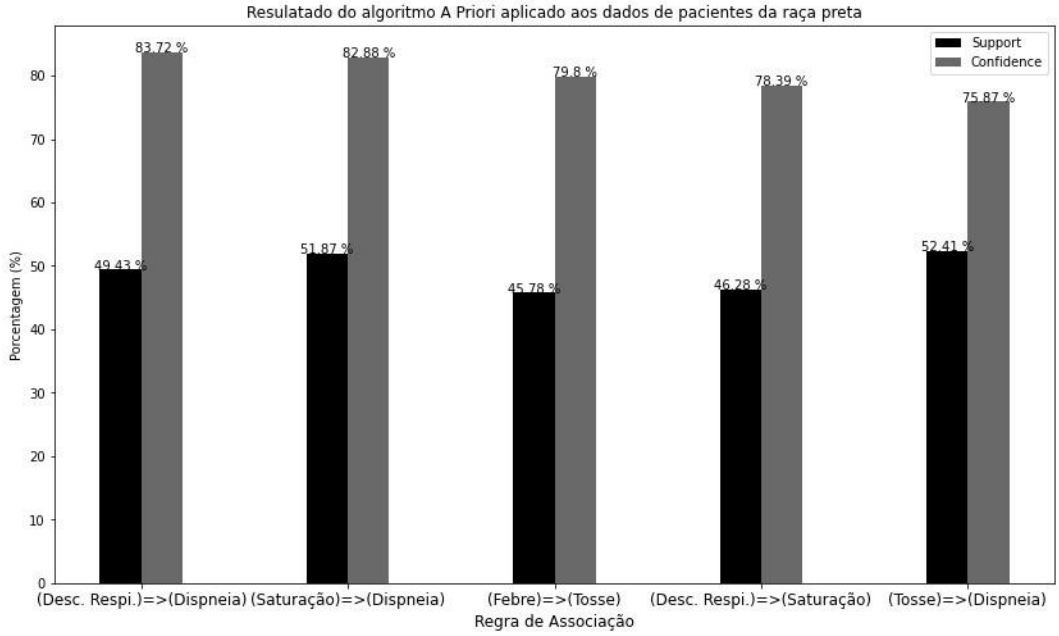


Figura 27 – **Resultado gráfico do Apriori relacionando sintomas dos pacientes da raça preta**

- 49,43% dos pacientes da raça preta que tiveram desconforto respiratório manifestaram também dispneia, sendo a confiança dessa regra acorrer de 83,72%;
- 51,87% dos pacientes da raça preta que tiveram saturação manifestaram também dispneia, sendo a confiança dessa regra acorrer de 82,88%;
- 45,78% dos pacientes da raça preta que tiveram febre manifestaram tosse, sendo a confiança dessa regra acorrer de 79,8%;
- 46,28% dos pacientes da raça preta que tiveram desconforto respiratório manifestaram saturação, sendo a confiança dessa regra acorrer de 78,39%;
- 52,41% dos pacientes da raça preta que tiveram tosse manifestaram também dispneia, sendo a confiança dessa regra acorrer de 75,87%;

Entre pacientes pretos a associação de sintomas que mais ocorre entre os casos é a de que se tiver a pessoa tiver saturação terá consequentemente dispneia, essa regra ocorre em cerca de 51,87% dos casos. Conferindo a confiança podemos afirmar que a maior probabilidade de uma regra ser verdadeira e a que se a pessoa preta tiver desconforto respiratório terá dispneia.

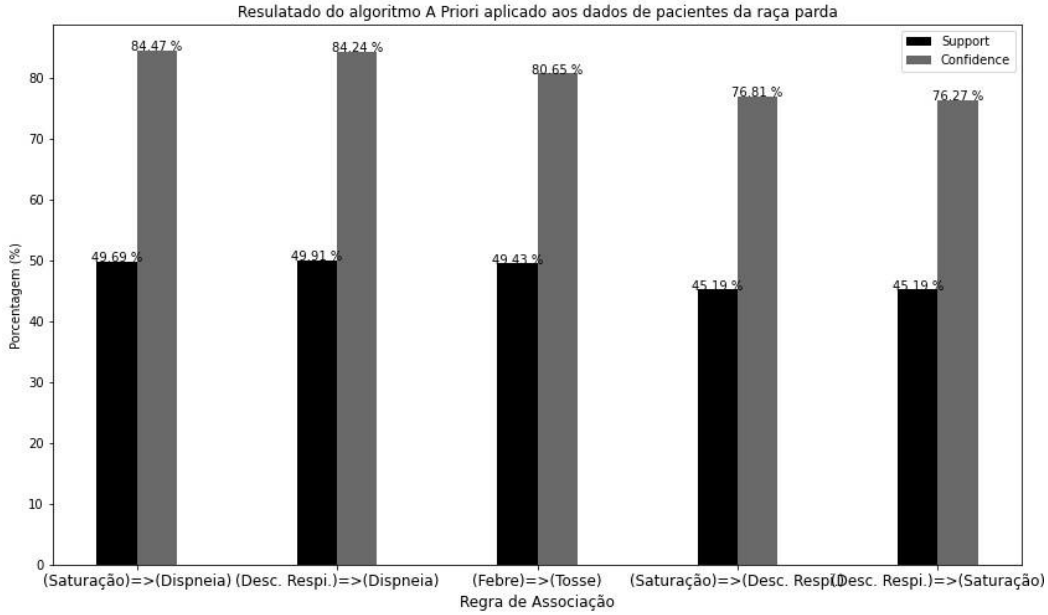


Figura 28 – **Resultado gráfico do Apriori relacionando sintomas dos pacientes da raça parda**

- 49,69% dos pacientes da raça parda que tiveram saturação manifestaram também dispneia, sendo a confiança dessa regra acorrer de 84,47%;
- 49,81% dos pacientes da raça parda que tiveram desconforto respiratória manifestaram também dispneia, sendo a confiança dessa regra acorrer de 84,24%;
- 49,43% dos pacientes da raça parda que tiveram febre manifestaram tosse, sendo a confiança dessa regra acorrer de 80,65%;
- 45,19% dos pacientes da raça parda que tiveram saturação manifestaram desconforto respiratório, sendo a confiança dessa regra acorrer de 76,81%;
- 45,19% dos pacientes da raça preta que tiveram desconforto respiratório manifestaram também saturação, sendo a confiança dessa regra acorrer de 76,27%;

Nos pacientes pardos a associação de sintomas que mais ocorre entre os casos a que se a pessoa tiver desconforto respiratório terá dispneia, essa regra ocorre em cerca de 49,91% dos casos. Observando a confiança podemos afirmar que a maior probabilidade de uma regra ser verdadeira entre os pardos e é a de que se a pessoa tiver saturação terá consequentemente dispneia.

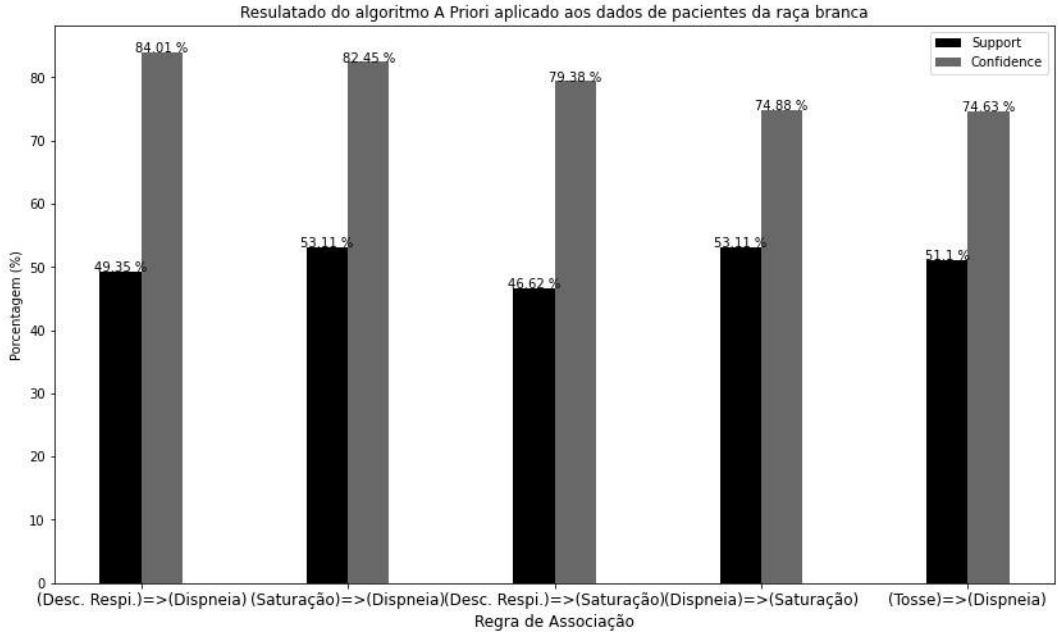


Figura 29 – **Resultado gráfico do Apriori relacionando sintomas dos pacientes da raça branca**

- 49,35% dos pacientes da raça branca que tiveram desconforto respiratória manifestaram também dispneia, sendo a confiança dessa regra acorrer de 84,01%;
- 53,11% dos pacientes da raça branca que tiveram saturação manifestaram também dispneia, sendo a confiança dessa regra acorrer de 82,45%;
- 46,62% dos pacientes da raça branca que tiveram desconforto respiratória manifestaram saturação, sendo a confiança dessa regra acorrer de 79,38%;
- 53,11% dos pacientes da raça branca que tiveram dispneia manifestaram saturação, sendo a confiança dessa regra acorrer de 74,88%;
- 51,1% dos pacientes da raça branca que tiveram tosse manifestaram também dispneia, sendo a confiança dessa regra acorrer de 74,63%;

Temos que entre os pacientes brancos a associação de sintomas que mais ocorre entre os casos a que se a pessoa tiver dispneia terá saturação, sendo essa afirmação valida também para a associação ao contrario (saturação resultando dispneia).Logo com a confiança pode-se afirmar que a associação desconforto respiratório implica dispneia tem a maior probabilidade de ser uma regra verdadeira entre os brancos.

Sabe-se que o maior índice de casos ocorre entre pessoas brancas seguido de pessoas pardas, sendo que pessoas pretas tem um dos menores índices de ocorrência. Logo sobre a relação dos dados raciais e de sintomas em comum, vemos que em pardos e brancos saturação implicando dispneia é a associação os que mais ocorre, já em pretos tosse implicando dispneia é tem maior ocorrência.

Considerando a confiança concluímos que nos pacientes pretos e brancos a associação desconforto respiratório implica dispneia tem a maior probabilidade de ser uma regra verdadeira entre esses dados, e que em pardos é a associação saturação resultando ter dispneia a maior.

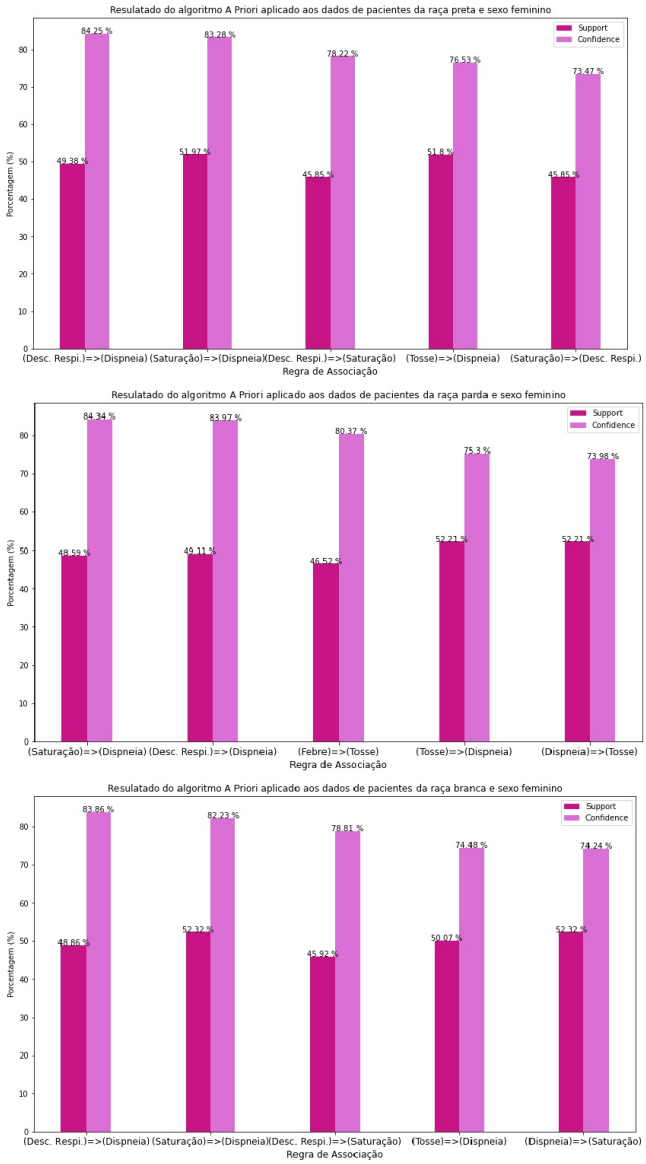


Figura 30 – Gráficos dos resultados do relacionamento do sexo feminino e raça dos pacientes de covid-19

O primeiro gráfico trás os pacientes do sexo feminino e da raça preta e suas associações de sintomas, vemos que a que mais aparece entre as regras de associação e a que saturação implica ter dispneia, com 51,97% entre os casos. Medindo a confiança, temos que desconforto respiratório resultando em ter dispneia como a regra com a maior probabilidade de ser verdadeira, com o valor de 84,25%.

No segundo mostra os pacientes do sexo feminino e da raça parda e suas regras de associações de sintomas, onde temos que o maior incidente é a regra tosse resultando dispneia e o contrario também, dispneia resultando tosse, ambos com 52,21%. Com a confiança temos

que saturação provocando dispneia no paciente como a associação com maior probabilidade de ser verdadeira.

Agora no terceiro e ultimo gráfico exibe-se os dados dos pacientes do sexo feminino e raça branca e as regras de associação de sintomas que essas características tem dentro da base, vemos que a regra dispneia resultando a pessoas ter saturação junto como o inverso (saturação implicando dispneia) como as como maior ocorrência entre os pacientes com essas caracterís-ticas, sendo de 52,32% . A confiança declara que desconforto respiratório provocando a pessoa com covid-19 ter dispneia é a associação com maior valor de ser verdadeira, 83,86%.

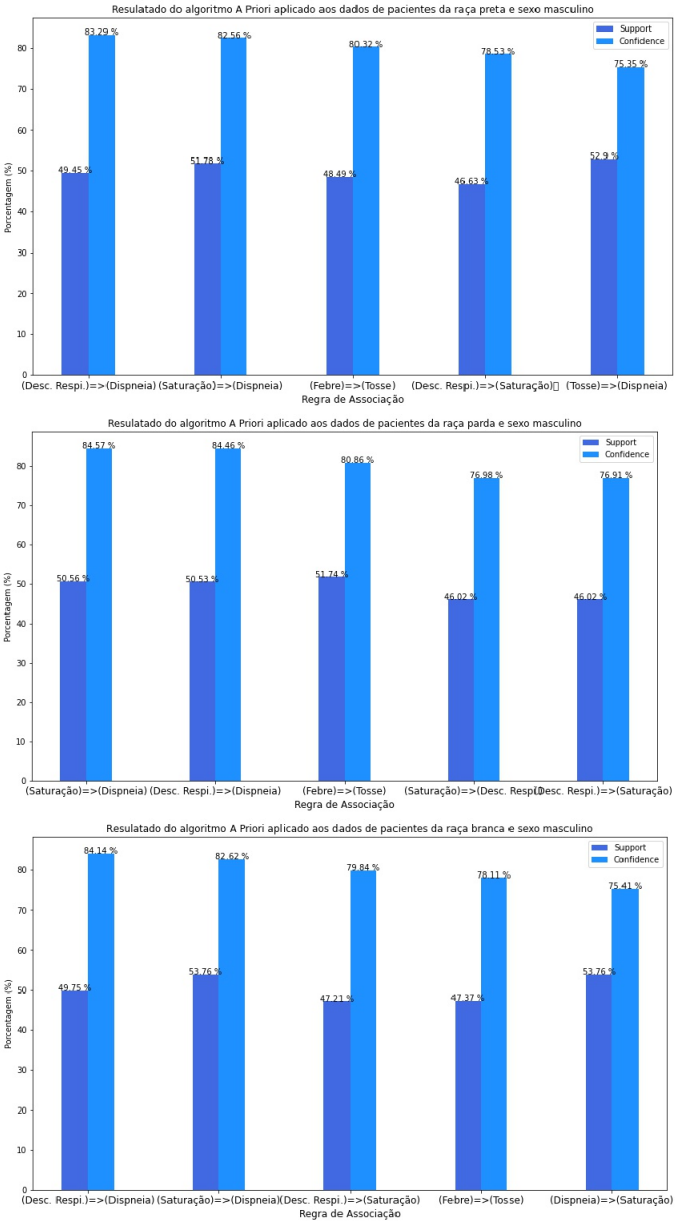


Figura 31 – Gráficos com os resultados do relacionamento entre o sexo masculino e a raça dos pacientes

Na figura 31 temos no primeiro gráfico os dados de pacientes do sexo masculino e raça preta e suas regras de associação de sintomas, temos como a maior ocorrência a regra

tosse implicando dispneia, sendo 52% de vezes que apareceu entre esses dados. Como maior confiança de ser uma regra verdadeira temos desconforto respiratório implicando dispneia, com 83,29%.

Segundo gráfico é dos pacientes do sexo masculino e raça parda com as regras de associação entre esse perfil, vemos que a regra com maior valor de incidência é febre resultando a pessoa ter tosse junto, como o numero de 51,74%. A confiança de ser uma regra verdadeira com maior valor temos saturação resultando a pessoa ter dispneia, com 84%.

No ultimo gráfico da figura 31 tem os pacientes do sexo masculino e raça branca e suas regras de associação de sintomas, a regra de maior ocorrência é saturação implicando o paciente ter dispneia, e vise e versa, como o valor de 53,76%. Como regra de maior confiança de ser verdadeira temos desconforto respiratório provocando o paciente ter dispneia também, com valor de 84,14%.

Por fim temos agora o Apriori aplicado as comorbidades, como foi dito anteriormente, o numero de casos com comorbidades tem um índice muito baixo comparado com os sintomas entre os dados, sendo assim foi optado por não aprofundar muito como nos sintomas.

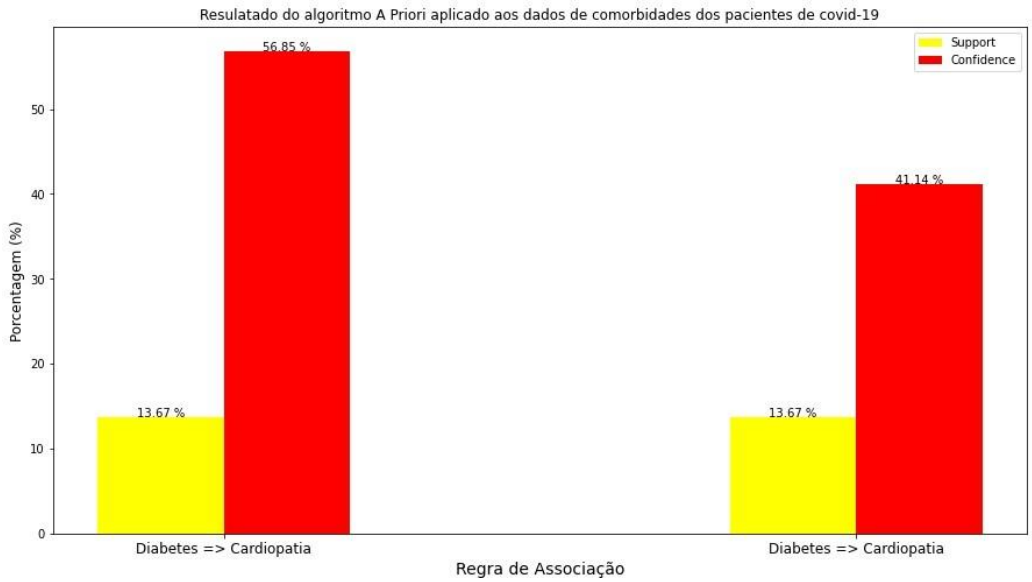


Figura 32 – Resultado gráfico do Apriori aplicado as comorbidades

Temos então na figura 32, que Diabetes resultando a pessoa ter cardiopatia é a regra de maior confiança e que ter diabetes e cardiopatia ao mesmo tempo ocorre em 13,67% dos casos. Conclui-se que diabetes e cardiopatia aparecem com mais frequência juntas entre os pacientes, assim ter essas doenças como comorbidade eleva o risco de manifestar mais severamente a covid-19.

Finalizando o projeto, concluímos então que conseguimos chegar em alguns conhecimentos interessantes a cerca dos dados, sendo que as comorbidades mais frequentes entre os

pacientes foram diabetes e cardiopatia e que são as que mais aparecem associadas entre os casos, também que sintomas de tosse, dispneia e desconforto respiratório são os que mais aparecem juntos entre os pacientes, e que a raça, sexo e idade influenciam nos sintomas e no numero de casos. Um ponto importante observado é que os resultados foram monótomos em alguns pontos, isso porque alguns sintomas são muito semelhantes, assim podendo causar certa repetição entre os resultados dificultando uma melhor visão do que se tem em cada ponto trabalhado.

REFERÊNCIAS

AGRAWAL, R.; IMIELIŃSKI, T.; SWAMI, A. Mining association rules between sets of items in large databases. In: **Proceedings of the 1993 ACM SIGMOD international conference on Management of data**. [S.l.: s.n.], 1993. p. 207–216.

CAMILO, C. O.; SILVA, J. C. d. Mineração de dados: Conceitos, tarefas, métodos e ferramentas. **Universidade Federal de Goiás (UFC)**, v. 1, n. 1, p. 1–29, 2009.

DABBURA, I. **K-means Clustering: Algorithm, Applications, Evaluation Methods, and Drawbacks**. 2018. Disponível em: <https://towardsdatascience.com/k-means-clustering-algorithm-applications-evaluation-methods-and-drawbacks-aa03e644b48a>.

DONGRE, J.; PRAJAPATI, G. L.; TOKEKAR, S. The role of apriori algorithm for finding the association rules in data mining. In: IEEE. **2014 International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT)**. [S.l.], 2014. p. 657–660.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. The kdd process for extracting useful knowledge from volumes of data. **Communications of the ACM**, ACM New York, NY, USA, v. 39, n. 11, p. 27–34, 1996.

HUREMOVIĆ, D. Brief history of pandemics (pandemics throughout history). In: **Psychiatry of pandemics**. [S.l.]: Springer, 2019. p. 7–35.

LATIF, S. et al. Leveraging data science to combat covid-19: A comprehensive review. **IEEE Transactions on Artificial Intelligence**, IEEE, v. 1, n. 1, p. 85–103, 2020.

LIU, Y. Study on application of apriori algorithm in data mining. In: IEEE. **2010 Second international conference on computer modeling and simulation**. [S.l.], 2010. v. 3, p. 111–114.

MAIMON, O.; ROKACH, L. Data mining and knowledge discovery handbook. Springer, 2005.

MINISTERIODASAUDE. **Open Data SUS**. 2020–2021.

OMS. **Coronavirus disease (COVID-19)**. 2021. Disponível em: <https://www.who.int/health-topics/coronavirus#tab=tab_3>.

PINHEIRO, A. M. e L. **OMS declara pandemia de coronavírus**. 2020. Disponível em: <https://g1.globo.com/bemestar/coronavirus/noticia/2020/03/11/oms-declara-pandemia-de-coronavirus.ghtml>.

TANDAN, M. et al. Discovering symptom patterns of covid-19 patients using association rule mining. **Computers in biology and medicine**, Elsevier, v. 131, p. 104249, 2021.

WERNECK, G. L.; CARVALHO, M. S. **A pandemia de COVID-19 no Brasil: crônica de uma crise sanitária anunciada**. [S.l.]: SciELO Public Health, 2020.