O PERFIL DA COVID-19 NO ESTADO DE SANTA CATARINA POR MEIO DE TAREFA DE CLASSIFICAÇÃO EM DATA MINING

Ana Claudia Fontana Medeiros[1], Merisandra Cortes de Mattos Garcia [2]

Resumo: Este artigo procura descrever o perfil da covid-19 no estado de Santa Catarina por meio da tarefa de classificação em data mining. As análises foram construídas com base no conjunto de dados do estado de Santa Catarina, além da construção de modelos capazes de prever os casos de recuperação e óbito no estado. O modelo do qual demonstrou uma acurácia de cerca de 98,95% de acerto.

Palavras-chave: Algoritmos de Classificação. *Data Mining*. Covid-19.

ABSTRACT: This paper has the objective to describe the profile of covid-19 in the state of Santa Catarina through the data mining classification task. The analyzes were built based on the data set from the state of Santa Catarina, in addition to the construction of models capable of predicting cases of recovery and death in the state. The model of which demonstrated an accuracy of about 98.95% of correctness.

Keywords: Classification Algorithms. *Data Mining*. Covid-19.

1 INTRODUÇÃO

O avanço na tecnologia de coleta e armazenamento de dados permitiu que organizações acumulassem uma vasta quantidade de dados. Muitas vezes, ferramentas e técnicas tradicionais de análise de dados não podem ser usadas devido ao tamanho do conjunto de informações (HAN; KAMBER, 2006; SCHEUNEMANN; PRETTO, 2015). Com a necessidade de estudar esses dados, surgiram novos métodos de busca para transformar dados em conhecimento útil (SUMATHI;

[1] Curso de ciência da Computação – Universidade do Extremo Sul Catarinense (UNESC) – Criciúma – SC- Brasil. anaclaudiaf.medeiros@gmail.com.
[2] Curso de ciência da Computação – Universidade do Extremo Sul Catarinense (UNESC) – Criciúma – SC- Brasil. mem@unesc.net.br.

SIVANANDAM, 2006, tradução nossa). Neste contexto, Fayyad em 1996 definiu o conceito de *data mining* como sendo a aplicação de algoritmos específicos de modo a identificar padrões a partir de dados existentes (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996, tradução nossa).

Data mining, segundo Tan, Steinbach e Kumar (2009) é o processo que busca descobrir de forma automática de conhecimento em base de dados a fim de identificar padrões e informações válidas, por meio de tarefas que envolvem estatística e aprendizado de máquina.

Por meio da classificação de dados, o *data mining* vem sendo utilizado em diversas áreas para que seja possível o auxílio de tomadas de decisões como nas áreas de saúde, bancárias, educação e outros.

Com o surgimento de novas doenças e pandemias em curso, a mineração de dados está sendo utilizada em projetos e pesquisas para que seja possível traçar o perfil de cada doença, pacientes além de fornecer auxílio para os profissionais da saúde nas tomadas de decisões, a exemplo do artigo de Albahri et al. (2020) que utiliza a tarefa de classificação em um conjunto de dados para que seja possível descrever o perfil de cada paciente além dos diagnósticos do novo coronavírus. Os resultados obtidos pela pesquisa podem ser utilizados por profissionais da saúde onde eles consigam ter um melhor controle e uma tomada de decisão melhor.

O novo coronavírus é uma doença infectocontagiosa que é causada pela síndrome respiratória aguda grave 2 (SARS-Cov2). O primeiro caso foi registrado na cidade de Wuhan, na China, em 31 de dezembro de 2019. Inicialmente, era tratado como uma pneumonia grave causada por um agente desconhecido pelas autoridades de saúde (ORGANIZAÇÃO MUNDIAL DA SAÚDE, 2019).

A tarefa de classificação é uma das técnicas mais usadas na mineração de dados e consistem em utilizar um conjunto de dados pré-classificados para que seja possível criar um modelo que faça a classificação de uma população de registros (DEULKAR; DESHMUKH, 2016).

Os modelos extraídos pelo classificador ajudam a fornecer uma melhor compreensão dos dados em geral, descrevendo o rótulo de classe, fazendo associação de um ou vários itens de dados nas classes que já estão predefinidas (DUDA; HART; STORK, 2001), esses modelos podem ser representados de várias

formas, como: árvore de decisão, regra de classificação, funções matemáticas e redes neurais (HAN; KAMBER; PEI, 2011).

A classificação é dividida em duas etapas, na primeira etapa o classificador é construído descrevendo um conjunto predeterminado de classes de dados que é chamado de etapa de aprendizagem. Na segunda etapa o modelo aprendido é usado para fazer a classificação nas tuplas de teste, a precisão do classificador é definida pela porcentagem das tuplas classificadas corretamente (BHATT; KANKANHALLI, 2011).

As tarefas de classificação contam com diversos algoritmos que permitem um grande ganho de desempenho em diferentes aplicativos.

Os algoritmos mais utilizados em classificação são os algoritmos de árvore de decisão C4.5, redes bayesianas *Naïve Bayes* e de redes neurais artificiais RBF, sendo eles reconhecidos pela comunidade científica dentre os dez algoritmos mais empregados na área (WU; KUMAR, 2009, tradução nossa). Esses algoritmos utilizam como entrada um conjunto de casos, cada um pertencendo a um pequeno número de classes e descrito pelos seus valores para um conjunto fixo de atributos (WU; KUMAR, 2009, tradução nossa).

O algoritmo C4.5 gera classificadores expressos em árvore de decisão, construindo classificadores em forma de conjunto de regras compreensíveis. Já o algoritmo *naïve bayes* tem fácil aplicação em um grande conjunto de dados, além de ser um algoritmo robusto e não tem a necessidade de aplicação de parâmetros iterativos (WU; KUMAR, 2009, tradução nossa), e por último o algoritmo RBF é utilizado na função de aproximação, ajuste de curvas e classificação de problemas. O RBF difere de outras redes neurais artificias que possuem características distintas devido a sua aproximação universal, topologia e velocidade de aprendizagem mais rápida (BESDOK; KURBAN, 2009, tradução nossa).

Considerando-se o exposto esta pesquisa, a partir do seu objetivo geral, propõe um modelo para identificação do perfil da covid-19 no estado de Santa Catarina por meio da classificação em *data mining* com a finalidade de identificar qual algoritmo de classificação obteve o melhor desempenho na tarefa de classificação. Tendo-se também como os objetivos: descrever o *data mining*, a tarefa de classificação e os algoritmos C4.5, *Naïve Bayes* e Radial Basis function; analisar os dados públicos referentes a Covid-19 no estado de Santa Catarina; aplicar a tarefa de

classificação por meio dos algoritmos C4.5, *Naïve Bayes* e RBF em dados da covid-19 e analisar por meio de medidas de qualidade em *data mining* o algoritmo que apresenta os melhores resultados para a base de dados do estudo.

Dentre os fatores que levaram a escolha de minerar os dados do novo coronavírus está o fato de que os dados serem públicos além na possibilidade que esses dados possam auxiliar profissionais da saúde a terem uma melhor decisão ao se deparar com os padrões de diagnósticos coletados, além de propor um melhor modelo para identificação do perfil da covid-19 no estado de Santa Catarina por meio da tarefa de classificação em *data mining*.

2 TRABALHOS CORRELATOS

Wignura e Riana (2020) propuseram a utilização do algoritmo C4.5 como auxílio no diagnóstico da covid-19 utilizando dados públicos da Indonésia. Os autores fazem a análise de diagnóstico do novo coronavírus por meio da tarefa de classificação em dados de pacientes confirmados, assintomáticos e pacientes que tiveram contato com casos confirmados ou sob suspeita, utilizando o algoritmo de árvore de decisão C4.5. Com os dados disponibilizados pelo governo da Indonésia, além do uso da metodologia de pesquisa experimental, que possibilita a identificação de fenômenos que permitem ao pesquisador manipular métodos para alcançar os objetivos, os atributos foram divididos em: febre, dificuldade respiratória, pneumonia grave e histórico de viagens para o exterior. Toda a análise foi realizada por meio do Rapdminer, onde o algoritmo C4.5 apresentou uma taxa de acurácia de 0,9286 (92,86%) de acerto.

Ibrahim et al. (2020) propõe a avaliação de desempenho dos algoritmos de redes neurais MLP e RBF para análise da disseminação da covid-19 e dos dados de óbito em 41 países da Ásia. Para a realização da análise do processamento, os dados foram divididos em dois conjuntos de treino e teste. Para o algoritmo MLP o conjunto de treino consiste em 78% dos dados gerais e para o RBF o conjunto de treino consiste em 73,2% dos dados gerais e o conjunto de teste consiste em 26,8% dos dados gerais. Os atributos utilizados para a realização dos dados foram os casos de óbitos, febre, população e porcentagem de mortes e casos totais. Ao final da pesquisa, constatou-se que o algoritmo RBF obteve um melhor desempenho em cima do

algoritmo MLP, sendo que o RBF teve uma taxa de acurácia de 0,498 e 0,003 apresentando uma taxa de erro mais baixa, já MLP apresentou uma taxa de acurácia de 0,498 e 0,002.

Muhammaed et al. (2020) apresenta a previsão da recuperação de pacientes infectados com covid-19 usando conjunto de dados epidemiológicos da Coréia do Sul. Para a realização da análise dos dados foi utilizada algoritmos de árvore de decisão, *N*aïve *Bayes*, SVM, Logistic Regression, Random Forest e K-Nearest utilizando a linguagem de programação Python para construir os modelos. Foram utilizados 1505 instancias de dados com 5 atributos úteis para o estudo, sendo eles o sexo, idade, caso de infecção e número de dias entre a data de confirmação e data de liberação ou óbito do paciente. Após a coleta de dados, a análise dos algoritmos foi feita utilizando as técnicas de acurácia, especificidade e sensibilidade. O resultado da pesquisa constatou que o algoritmo de árvore de decisão teve melhor desempenho em cima dos outros algoritmos, com uma taxa de acerto de 99,85%, seguido pelo Random Forest que obteve uma taxa de acerto de 99,60%, os demais algoritmos obtiveram taxas de acerto abaixo de 98%.

3 MATERIAIS E MÉTODOS

Esta pesquisa tem como objetivo escolher um algoritmo de classificação que prevê os casos de recuperados ou óbitos nos dados provenientes de casos da covid-19 no estado de Santa Catarina. Para isso, o conjunto de dados é analisado por meio de tarefa de classificação, utilizando os algoritmos de árvore de decisão, redes bayesianas e redes neurais artificiais para a geração de modelos. Os resultados obtidos são comparados por meio de medidas de qualidades, como a acurácia, Kappa, TP-Rate e F-Measure para identificar o algoritmo com melhor desempenho.

3.1 Tarefa de classificação

A tarefa de classificação tem o objetivo de descobrir funções que relacione um conjunto de registros com um determinado conjunto de classes, de modo que o processo de classificação possa usá-la para predizer a classe de um exemplo novo e desconhecido (HAN; KAMBER; PEI, 2011).

Para realizar a classificação são aplicados diversos algoritmos, onde destacam-se os algoritmos de redes neurais, classificadores bayesianos e árvores de decisão (FAYYAD; PITATETSKY-SHAPIRO; SMYTH, 1996; GOLDSCHMIT, PASSOS; BEZERRA, 2015).

3.2 Medidas de qualidade

Um dos pontos chaves no processo de data mining, a avaliação de modelos analisa o desempenho de cada algoritmo separadamente, pois cada algoritmo possui limitações, então torna-se importante a utilização de metodologias de avaliação que compare cada algoritmo utilizado durante o processo de data mining (TAN; STEINBACH; KUMAR, 2009).

As medidas de qualidades obtidas por meio da matriz de confusão, que são os registros previsto pelo conjunto de testes, são bases para calcular algumas das medidas de qualidades, as mais utilizadas são: acurácia, que faz a medida de eficiência geral do classificador; precisão, mede se a porcentagem de instancias classificadas como positivas estão corretas; F-measure, mede a média entre a precisão e a sensibilidade; AUC, faz a medida da capacidade do classificador desviar de classificação falsa; sensibilidade, mede a capacidade de classificação do modelo; ROC, mostra o quanto o modelo criado consegue distinguir entre dados binários (LAROSE, 2015, tradução nossa).

Todas as informações obtidas no processo de data mining têm a sua relevância avaliada por meio do desempenho do classificador. A qualidade de um classificador é medida por meio da taxa de erro, que corresponde ao número de classificação incorreta de um conjunto de dados (WITTEN; FRANK; HALL, 2011).

3.3 Covid-19 no Brasil

No Brasil o primeiro registro de covid-19 foi, em 26 de fevereiro de 2020, no estado de São Paulo. Até o mês de novembro de 2021 o número de casos confirmados no Brasil chegou a 21.939.196 e de óbitos 610.491 (BRASIL, 2021).

A infecção pela covid-19 ocorre de pessoa para pessoa por meio de gotículas transportadas pelo ar, sendo elas originadas por tosse ou espirro de uma pessoa

contaminada, ou por meio do toque em uma superfície contaminada e levar a mão a boca, nariz ou olhos (TESINI, 2020).

A melhor forma de conter a propagação do vírus é por meio de isolamento e distanciamento social, apesar do avanço das vacinações em todo território nacional, é importante que seja mantido o uso de máscaras e isolamento social, a fim de que a vacinação atinja uma parte considerável da população e vá diminuindo gradualmente (BRASIL, 2021).

3.4 Base de dados

A base de dados utilizada nesta pesquisa é disponibilizada pelo portal de dados abertos do estado de Santa Catarina[3] em formato de arquivo cvs. Os dados são organizados pelos casos positivos que ocorreram no estado de Santa Catarina até o dia 15 de agosto de 2021. Ao todo a base conta 1.048.576 registros.

A tabela 2, descreve cada campo que compõe a base disponibilizada pelo portal boa vista.

Tabela 2 – dicionário de dados

Atributo	Tipo	Descrição
data_publicacao	Texto	Data de publicação do conjunto de dados no portal de dados abertos
recuperados	Texto	Indicação de que o paciente foi recuperado
data_inicio_sintomas	Texto	Data do início dos sintomas
data_coleta	Texto	Data da coleta da amostra
estado	Texto	Sintomas do paciente
comorbidades	Texto	Comorbidades do paciente
gestante	Texto	Indica os casos de gestantes ou puérpera
internacao	Texto	Indicação de que o paciente está internado
internacao_uti	Texto	Indicação de que o paciente está internado em UTI
sexo	Texto	Indicação de sexo biológico do paciente
municipio	Texto	Município de residência do paciente
obito	Texto	Indicação de que o paciente veio a óbito
data_obito	Texto	Data do óbito do paciente
idade	Numérico	Idade do paciente
regional	Texto	Mesorregião de residência do paciente
raca	Texto	Raça do paciente
data_resultado	Timestamp	Data e hora da confirmação
codigo_ibge_municipio	Numérico	Código do IBGE do município de residência do paciente

[3] Dados disponibilizados em: http://dados.sc.gov.br/ca/dataset?tags=COVID-19

latitude	Numérico	Latitude do município de residência do paciente
longitude	Numérico	Longitude do município de residência do paciente
estado	Texto	Nome do estado de residência do paciente
criterio_confirmacao	Texto	Critério utilizado para confirmação do caso
tipo_teste	Texto	Tipo de teste utilizado para confirmação do caso
municipio_notificacao	Texto	Município onde a notificação foi registrada
codigo_ibge_municipio_notificacao	Numérico	Código do IBGE do município onde a notificação foi registrada
latitude_notificacao	Numérico	Latitude do município onde foi registrada a notificação
longitude_notificacao	Numérico	Longitude do município onde foi registrada a notificação
classificacao	Texto	Classificação de confirmação de caso positivo
origem_esus	Texto	Indica que a origem da informação se encontra no e-SUS VE
origem_sivep	Texto	Indica que a origem da informação se encontra no SIVEP Gripe
origem_lacen	Texto	Indica que a origem da informação se encontra no LACEN/SC
origem_laboratorio_privado	Texto	Indica que a origem da informação se encontra em um Laboratório Privado
nome_laboratorio	Texto	Quando campo origem_laboratorio_privado for preenchido, neste campo constará o nome do laboratório
fez_teste_rapido	Texto	Indicativo se o paciente (sem informação do paciente) fez teste rápido
fez_pcr	Texto	Indicativo se o paciente (sem informação do paciente) fez PCR
data_internacao	Texto	Data da internação informada no SIVEP Gripe
data_entrada_uti	Texto	Data da internação UTI informada no SIVEP Gripe
regional_saude	Texto	Data da evolução do caso informada no SIVEP Gripe
data_evolucao_caso	Texto	Data da internação informada no SIVEP Gripe
data_saida_uti	Texto	Regional de saúde do município de notificação
bairro	Texto	Bairro do Paciente

Fonte: Santa Catarina (2021).

3.5 Pré-processamento

O pré-processamento é a etapa que prepara o conjunto de dados para que seja possível a realização da mineração de dados. Para realizar o pré-processamento do conjunto de dados, foram necessárias ferramentas como Excel 365 e WEKA 3.8.5.

3.5.1 Seleção de atributos

Os seguintes atributos foram selecionados: recuperados, gestante, internação, internação uti, sexo, município, óbito, regional, critério de confirmação, tipo de teste, município de notificação, origem esus, origem sivep, origem lacen, origem laboratório privado, fez teste rápido, fez pcr, regional saúde, idade, sintomas,

comorbidades e desfecho (classe criada pelos atributos referentes a recuperado e óbito). Os campos de sintomas e comorbidades foram divididos conforme os registros existentes na coluna.

Todos os registros dos atributos recuperados e óbitos foram analisados para que fossem retirados os registros onde se encontravam “não” para recuperados e “não” para óbito e onde os registros se encontravam com “sim” para recuperados e “sim” para óbito, todos os campos correspondentes aos sintomas e comorbidades que estavam vazios foram considerados como sem sintomas e sem comorbidades.

Após limpeza dos dados e divisão dos valores, ao todo foram contabilizados 1.037.436 registro.

Ao todo 1.020.052 registros foram apresentados como recuperados e 17.384 dos casos evoluíram para óbito.

3.5.2 Balanceamento de classes

A falta de balanceamento de classe pode influenciar o desempenho dos algoritmos de classificação, pois esses algoritmos são sensíveis ao desbalanceamento fazendo com que elas levem em consideração as classes predominantes ignorando as classes de menor proporção (WEISS, 2004).

Como a classe de desfecho encontrava-se desbalanceada, onde possui 1.020.052 registros para “recuperados” e 17.384 para “óbito”. Sendo a diferença desses registros de 1.002.668, sendo a classe de “recuperados” maior, partindo do ponto em que o número de óbitos é menor que os recuperados.

A alternativa para que sejam definidos pesos ou tamanhos similares para os registros foi a aplicação do filtro supervisionado SMOTE. A técnica SMOTE faz a adição de novos casos para a classe, isto é, gera casos sintéticos para a classe minoritária a partir dos casos já existentes (CHAWLA et al, 2002).

O filtro supervisionado SMOTE foi aplicado com o parâmetro de 100% e o número de vizinhos com o valor 5, conforme a ferramenta WEKA sugere. Ao fazer a aplicação do filtro, obteve-se um aumento dos registros de óbito, que passou de 17.384 registros para 34.768, sendo metade desses registros dados sintéticos. A diferença entre os registros de recuperados e óbito passou a ser de 967.900 registros.

3.5.3 Discretização dos dados

A tarefa de discretização é importante para os algoritmos de classificação, pois é nela que são realizadas as separações dos dados com uma relevância maior, convertendo um atributo contínuo em atributo discreto, definindo as categorias e mapeamento de valores (TAN; STEINBACH; KUMAR, 2009).

O método de discretização foi aplicado nos registros de idade, onde foi transformado em faixa etária colocando-se as idades em intervalos de 10 anos conforme estipulado pela Organização das Nações Unidas (ONU) e Organização Mundial da Saúde (OMS).

3.5.4 Normalização

A tarefa de normalização de dados faz a transformação dos intervalos originais dos registros em um intervalo específico. A utilização da normalização dos dados é valiosa para os métodos que fazem cálculo de distância entre os atributos, como por exemplo, o método que faz a utilização como o “k-vizinhos mais próximos” costuma a dar mais importância aos atributos que tem um intervalo maior em seus valores (TAN; STEINBACH; KUMAR, 2009). A ferramenta WEKA disponibiliza a técnica de normalização como um filtro não supervisionado e a aplicação dessa técnica foi utilizada em apenas alguns experimentos feitos nesta pesquisa.

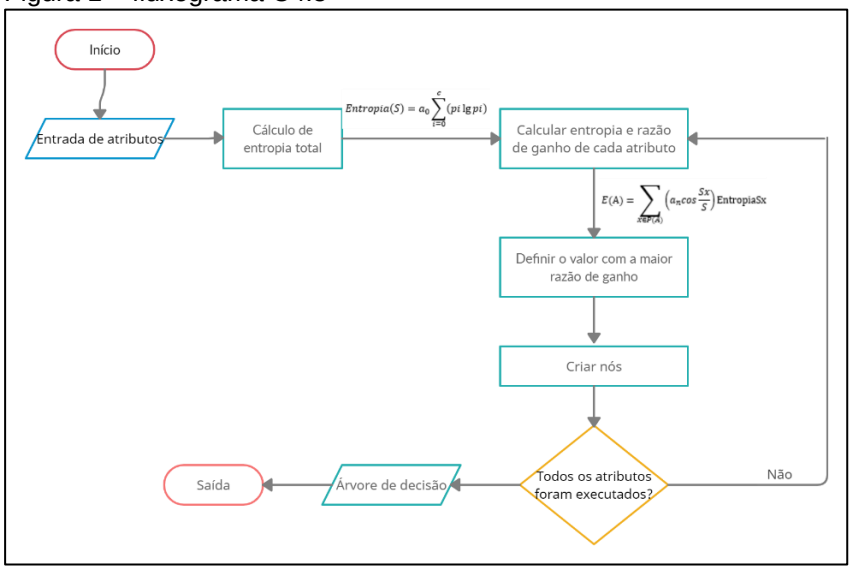
3.6 Experimentos

Nesta pesquisa foram realizados quatro experimentos, utilizando os algoritmos de classificação: C4.5 que é um classificador de árvore de decisão, onde são geradas árvores de decisão podadas ou não; RBF utilizado para resoluções de problemas de aprendizado de máquina e *Naïve Bayes* que calcula um conjunto de probabilidades contando a frequência e as combinações de valores em uns determinados conjuntos de dados; na tentativa de obter um melhor resultado na previsão dos casos de óbito. A escolha dos algoritmos de classificação se deu pelo fato de serem algoritmos amplamente utilizados em pesquisas, além de estarem entre os dez algoritmos mais utilizados para classificação.

3.6.1 Algoritmo C4.5

Amplamente utilizado no aprendizado de máquina e para indução de árvores de decisão, o algoritmo C4.5 gera um classificador em que a estrutura criada possui dois tipos de nós, onde uma folha indica uma classe ou um nó de decisão que irá especificar os testes que serão executados em um valor de atributo único, com um ramo e uma subárvore para cada resultado possível do teste (KANTARDZIC, 2011, tradução nossa). O classificador criado faz a divisão dos dados de forma recursiva a partir de dados de treino. Cada conjunto de dados possuí uma árvore construída com base nos atributos que possuem um maior ganho de informações (WU; KUMAR, 2009, tradução nossa). A figura 1 apresenta o fluxograma do algoritmo C4.5.

Figura 1 – fluxograma C4.5



Fonte: Adaptado de Anwar, Pranolo e Kurnaiwan (2017, tradução nossa).

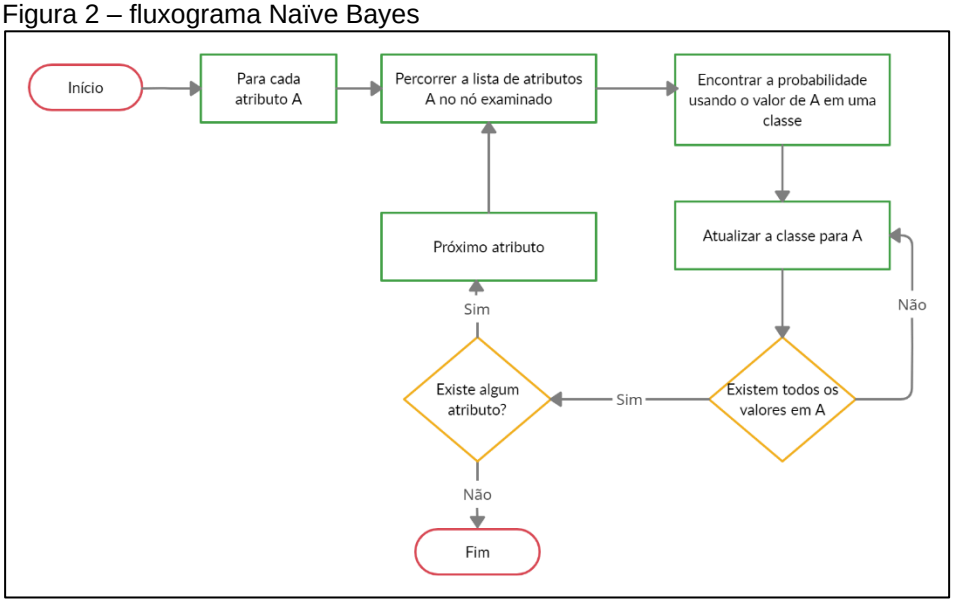
3.6.2 Algoritmo Naïve Bayes

Baseado no teorema de Bayes, o classificador bayesiano utiliza o ramo da matemática, conhecido como teoria da probabilidade para identificar a classificação mais provável em dados (LAROSE, 2005; KAMAT, 2009, tradução nossa).

O classificador bayesiano é um classificador estatístico que faz previsões das probabilidades de associação em uma determinada classe. O algoritmo *Naïve Bayes* é um classificador probabilístico simples, que calcula um conjunto de probabilidades contando a frequência e as combinações de valores em uns

determinados conjuntos de dados, figura 2 (PATIL; SHEREKAR, 2013, tradução nossa). Devido a sua simplicidade o *Naïve Bayes* estima os parâmetros do modelo a partir de um conjunto de dados usando a estatística bidimensional para a classe e cada atributo, tornando o processo de classificação eficiente (ABELLÁN; CASTELLANO, 2017, tradução nossa).

Figura 2 – fluxograma Naïve Bayes



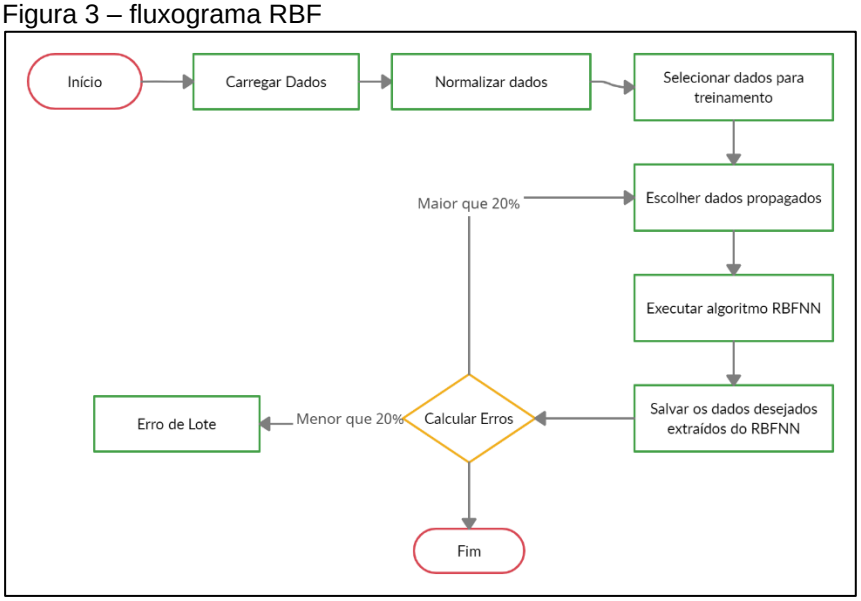
Fonte: Adaptado de Sneha e Gangil (2019, tradução nossa).

3.6.3 Algoritmo Radial Basis Function

O *Radial Basis Function* (RBF) é um algoritmo de redes neurais artificiais empregado para a resolução de problemas de classificação. As unidades ocultas em uma RBF oferecem um conjunto de funções que são como uma base para os padrões de entrada, quando eles são colocados sobre o espaço oculto, essas funções são conhecidas como *Radial Basis Function* (Funções de base radial). Cada coordenada irá caracterizar-se por definir o centro de uma região que possui uma maior aglomeração do espaço de dados de entrada.

Segundo Haykin (2001) a construção de uma RBF envolve a camada de entrada constituída por nós que conectam à rede ao seu ambiente; a camada oculta que aplica uma transformação não-linear do espaço de entrada para o espaço oculto; e a camada de saída que fornece a resposta da rede ao padrão de ativação aplicado a camada de entrada.

Figura 3 – fluxograma RBF



Fonte: Adaptado de Mamat et al. (2020, tradução nossa).

3.6.4 Execução dos experimentos

Para realizar os experimentos, foi feito o pré-processamento de dados em cima do conjunto de dados, disponibilizado pelo estado de Santa Catarina, da covid-19, ao todo foram 1.037.436 registros classificados.

A execução desses experimentos foi feita por meio do uso da validação cruzada ou *Cross-Validation* com o valor de iterações estabelecidos sendo igual à dez. O uso das iterações da validação cruzada sendo igual à dez permite que seja possível obter a acurácia ou *F-Measure* maior.

A execução do *data mining* foi realizada na ferramenta *WEKA*, que possuí uma vasta literatura sobre seu funcionamento de utilização além de contar com os algoritmos de classificação que foram utilizados nessa pesquisa. O *WEKA* disponibiliza também algoritmos de fragmentação que trata a forma de como o programa irá interagir com a base a ser utilizada, os dois disponibilizados são o *k-folds cross-validation* e *percentage split*.

Casos com uma base de dados muito grande para validação ela poderá ser dividida em duas partes, sendo elas conjunto de treino e conjunto de testes.

O método de *k-folds cross-validation* é utilizado para mitigar problemas de conjuntos de dados não muito grandes para ser dividido em tantas partes. No

procedimento é feito a escolha de número de *folds*, que fará a divisão dos dados em n partes, por exemplo, se for considerado o número de *folds* em quatro, os dados serão divididos em quatro partes (WITTEN et al., 2011).

Nos experimentos realizados o número de *folds* utilizado foi de dez partições, conforme a literatura pede, pois esse número é o mais adequado para que se obtenha uma estimativa de erro ideal (WITTEN et al., 2011).

Para a execução dos algoritmos de classificação foram definidos quatro experimentos, os quais serão descritos na tabela 2. O objetivo desses experimentos é analisar o desempenho de classificação dos algoritmos de classificação no conjunto de dados.

Tabela 2 – experimentos realizados

Experimento	Descrição
1	Dados discretizados, normalizados e com balanceamento de classe, utilizando o desfecho como classe
2	Dados discretizados, normalizados e com balanceamento de classes utilizando recuperados como classe
3	Dados discretizados, normalizados e com balanceamento de classe, utilizando óbito como classe
4	Dados discretizados, normalizados e sem balanceamento de classe, utilizando desfecho como classe

Fonte: Do autor.

Como a classe desfecho encontrava-se desbalanceada, em alguns experimentos, foi necessário a aplicação do filtro SMOTE que permite a criação de dados sintéticos a fim de que os dados fiquem igualados.

Os resultados obtidos por meio dos algoritmos de classificação são analisados por meio das medidas de qualidade.

3.7 Análise de resultados

Para Tan, Steinbach e Kumar (2009) as métricas principais para avaliar modelos são acurácia, TP-Rate, matriz de confusão, coeficiente Kappa e curva ROC. Ainda de acordo com Tan, Steinbach e Kumar (2009) para realizar uma avaliação geral do desempenho do modelo é possível utilizar as medidas como kappa e acurácia, porém quando o conjunto de dados possui classes desbalanceadas, as medidas como acurácia pode não ser o mais indicado na avaliação do modelo.

Nesta pesquisa foram utilizadas as medidas de qualidade como *F-Measure*, para analisar a razão das somas dos acertos totais, acurácia, coeficiente Kappa, TP-Rate para identificação dos verdadeiros positivos.

A análise da significância estatística dos percentuais obtidos na acurácia foi realizada por meio do teste estatístico *T-Test*, disponível no WEKA na aba experimenter. O nível de significância utilizado para avaliar a acurácia foi de 5%, conforme a própria ferramenta sugere.

4 RESULTADOS E DISCUSSÃO

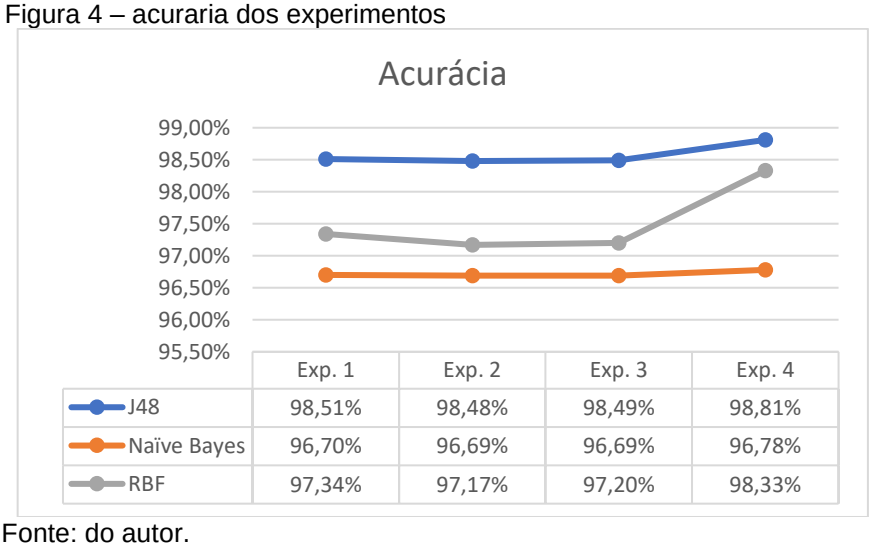
Considerando-se os experimentos realizados, os resultados foram analisados por meio dos percentuais de acurácia, coeficiente Kappa, TP-Rate e F-Measure, que foram gerados por cada algoritmo. O intuito dessas análises é identificar qual experimento gerou os melhores modelos em cima do conjunto de dados de Santa Catarina até o dia 15 de agosto de 2021.

4.1 Resultados gerados pelos experimentos

Os modelos obtidos pelos algoritmos de classificação, foram analisados por meio de medidas de qualidade em classificação.

A figura 4 apresenta a acuraria atingida pelos classificadores nos quatro experimentos realizados.

Figura 4 – acuraria dos experimentos



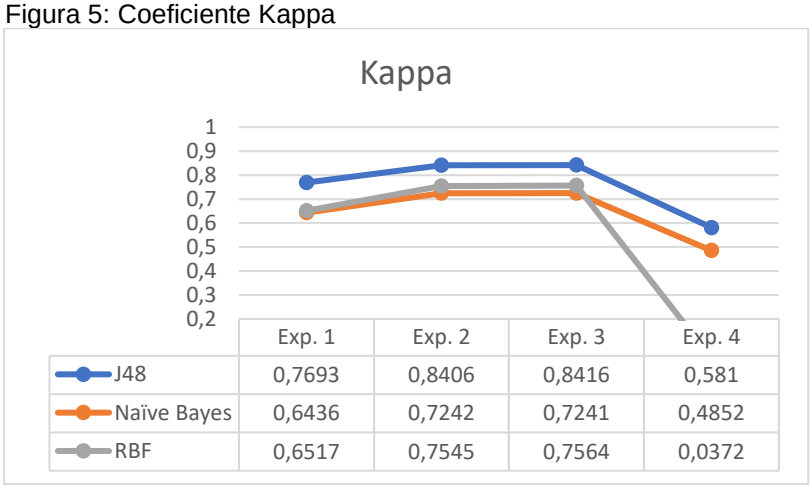
	Exp. 1	Exp. 2	Exp. 3	Exp. 4
J48	98,51%	98,48%	98,49%	98,81%
Naïve Bayes	96,70%	96,69%	96,69%	96,78%
RBF	97,34%	97,17%	97,20%	98,33%

Fonte: do autor.

Na implementação dos modelos de classificação em cima dos casos positivos no estado de Santa Cataria, o melhor resultado foi obtido utilizando o algoritmo de árvore de decisão C4.5 junto com a aplicação do filtro de balanceamento de classes SMOTE no experimento 1 utilizando o atributo desfecho como classe. O algoritmo C4.5 obteve o valor de acurácia de 98,95% no primeiro experimento. Em seguida o algoritmo de redes neurais RBF obteve o valor de acurácia de 97,34% e o algoritmo de redes bayesianas obteve o valor de 96,70%.

Para o coeficiente Kappa os algoritmos que tiveram o melhor resultado são os que possuem o valor mais próximo de 1, conforme a figura 5 apresenta.

Figura 5: Coeficiente Kappa



	Exp. 1	Exp. 2	Exp. 3	Exp. 4
J48	0,7693	0,8406	0,8416	0,581
Naïve Bayes	0,6436	0,7242	0,7241	0,4852
RBF	0,6517	0,7545	0,7564	0,0372

Fonte: do autor.

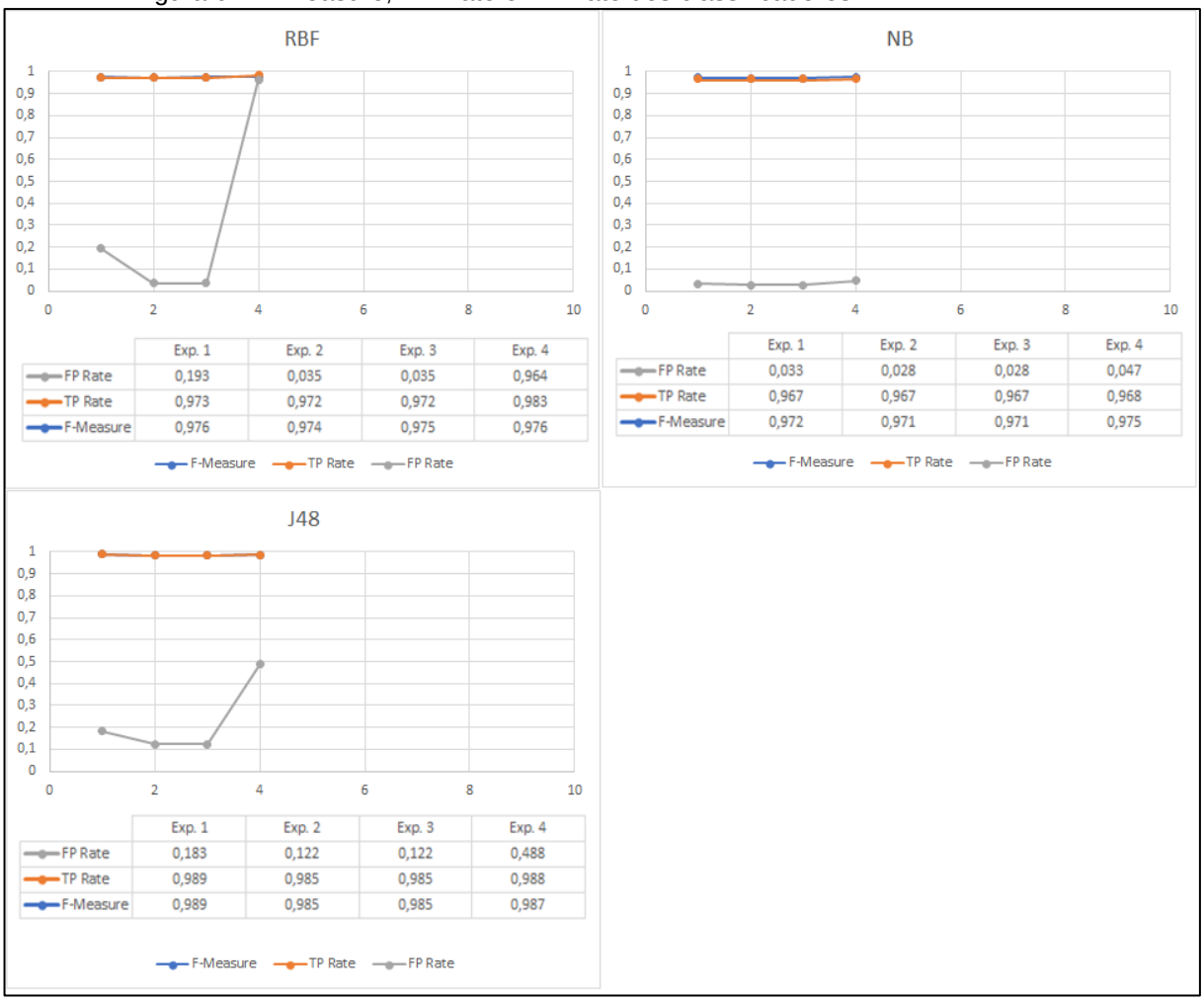
Foi observado na figura 5, o algoritmo de classificação C4.5 obteve o maior índice com 0,832 de acurácia no experimento 1, seguido pelo algoritmo de redes neurais RBF que obteve 0,6517 e do algoritmo de redes bayesianas *Naïve Bayes* que obteve 0,6436 de acurácia.

Nota-se também a queda de valor do kappa no experimento 4 no algoritmo RBF, essa queda ocorre devido ao desbalanceamento de classes que existem nos dados pré-processados.

Os índices mais altos foram nos experimentos que fizeram a utilização do balanceamento de classe por meio da aplicação do filtro SMOTE.

Na figura 6, pode-se observar o valor do F-Measure que obteve valores consideráveis tanto para os classificadores positivos quanto aos negativos.

Figura 6 – F-Measure, TP Rate e FP Rate dos classificadores



Fonte: do autor.

O *F-Measure* obteve um valor considerável próximo do 100% de relevância para os classificadores positivos tanto para os classificadores negativos. No experimento 1, classe desfecho, o algoritmo C4.5 obteve 0,989 de classificação, seguido pelo algoritmo RBF que obteve uma taxa de 0,976 e o NB 0,972. Um dos motivos para a classificação ter obtido um valor acima de 95%, no experimento 1, foi a utilização do filtro SMOTE que fez o balanceamento da classe.

Para verificar a taxa de acerto dos algoritmos em relação aos experimentos realizados, foi utilizado a técnica de teste estatístico T, disponível na guia explorer do WEKA. O experimento foi utilizado a fim de testar a significância estatística dos resultados obtido no experimento 1, que obteve um melhor desempenho em comparação aos outros experimentos, com o nível de significância em 5%, conforme a ferramenta sugere.

Tabela 5 – Taxa de acerto dos algoritmos

Algoritmos	C4.5	NB	RBF
Acurácia	98.51	96.70	97.34
Desvio Padrão	0.04	0.05	0.11
Kappa	0.77	0.64	0.65
Marcação		*	*

Fonte: do autor.

A tabela 5 mostra os resultados obtidos do desvio padrão e da acurácia. Caso um classificador obtenha um resultado abaixo do esperado pelo teste é feita a marcação do mesmo com um asterisco (*), caso o algoritmo tiver um melhor resultado a marcação é feita com um (v).

Vale observar que o algoritmo C4.5 obteve um percentual de acurácia maior que os demais algoritmos, sendo ele de 98,51%. Os algoritmos RBF e NB possuem uma diferença estatística maior para o nível de significância considerado de 5%.

Tabela 6 – F-measure e TP Rate

Algoritmos	C4.5	NB	RBF
TP Rate	0.99	0.97 *	0.98 *
F Measure	0.99	0.98 *	0.99 *

Fonte: do autor.

Novamente vale observar que os valores dos verdadeiros positivos são maiores para o algoritmo C4.5 com 0.99, já o algoritmo RBF possui uma taxa de 0,98 e o NB com 0,97.

Os valores do *F-Measure* possuem uma maior taxa para os algoritmos RBF e C4.5, ambos com 0,99, logo atrás o algoritmo NB possui uma taxa de 0,98.

Vale observar que o teste de significância aplicado nos percentuais de acurácia e Kappa mostrou que o C4.5 possui um melhor modelo de classificação quando comparado aos outros dois algoritmos.

4.2 Discussão dos resultados

Em relação aos experimentos realizados, pode-se observar que os algoritmos que não possuem balanceamento de classe, possuem taxas de verdadeiro positivos altas. Os valores dos algoritmos de classificação obtiveram valores de acerto próximos aos trabalhos propostos, a exemplo do algoritmo C4.5 que obteve uma taxa

de acurácia de 98,95%, que se aproxima do trabalho proposto por Wignura e Riana (2020). O algoritmo RBF obteve um valor de acurácia muito próximo do trabalho proposto por Ibrahim et al. (2020), sendo o valor obtido por esta pesquisa de 97,34% para 0,498 da pesquisa proposta por Ibrahim.

5 CONCLUSÃO

Apesar de ser uma doença sem medicações eficazes, a Covid-19 deve ser evitada ao máximo. Buscando registrar algumas condições sobre as características mais presentes em pacientes que tiveram recuperação ou foram a óbito, dados capturados até do dia 15 de agosto de 2021 no estado de Santa Catarina foram utilizados para a construção de um modelo onde seja possível prever os óbitos em Santa Catarina. O objetivo geral do trabalho em propor um modelo para identificação do perfil da Covid-19 no estado de Santa Catarina por meio da tarefa de classificação em *data mining* foi atingido, onde os resultados apresentados mostraram que o algoritmo de árvore de decisão C4.5 obteve o melhor desempenho em todos os experimentos executados. O modelo elaborado foi utilizado para auxiliar na previsão dos óbitos e dos recuperados no estado de Santa Catarina, a previsibilidade do modelo se mostrou satisfatória pois atendeu a uma acurácia média e para os casos positivos maior que 80%, considerando os dados utilizados com registro até 15 de agosto de 2021. Apesar das dificuldades, os resultados encontrados são satisfatórios e atingem os objetivos da pesquisa. Após análises comparativas entre os algoritmos de classificação, valida-se que a acurácia de um modelo pode ter grandes variações quando utilizado balanceamento de classes.

Considerando os resultados obtidos nesta pesquisa, destacam-se algumas sugestões de trabalhos futuros:

a) Utilizar conjunto de dados em todo território nacional a fim de traçar o perfil de recuperados e dos que tiveram óbito em todo o país;
b) Aplicar outros algoritmos de classificação para análise dos modelos gerados;
c) Analisar os resultados por meio de outras medidas de qualidade em data mining;
d) Analisar a sensibilidade e especificidade dos algoritmos utilizados.

REFERÊNCIAS

A_comparative_study_of_decision_tree_ID3_and_C4.5.pdf>. Acesso em: 13 nov. 2021.

BRASIL. **BOLETIM EPIDEMIOLOGICO**. 2021. Disponível em: <https://www.gov.br/saude/pt-br/media/pdf/2021/agosto/20/boletim_epidemiologico_covid_76-final20ago.pdf> Acesso em: 01 nov. 2021.

DEULKAR, Miss Deepa S; DESHMUKH, R. R. **Data Mining Classification**. Disponível em:<http://www.imperialjournals.com/index.php/IJIR/article/view/299>. Acesso em:10 out. 2021.

FAYYAD, USAMA; PIATETSKY-SHAPIRO, GREGORY; SMYTH, PADHRAIC, **From Data Mining to Knowledge Discovery in Databases**, AI Magazine Volume, 17 Number 3, 1996.

GOLDSCHMIDT, RONALDO; PASSOS, EMMANUEL; BEZERRA, EDUARDO. **Data Mining: Conceitos, técnicas, algoritmos, orientações e aplicações**, 2 ed. Rio de Janeiro: Elsevier, 2015.

HAN, Jiawei; KAMBER, Micheline. **Data mining: concepts and techniques**. San Francisco:Morgan Kaufmann, 2001.

HAN, J.; KAMBER, M. **Data Mining: Concepts and Techniques**.California, 2.ed. Morgan Kaufmann, 2006.

Han, Jiawei; Kamber, Micheline; Pei, Jian. **Data mining: concepts and techniques 3rd ed**. Massachusetts: Elsevier, 2012.

HSSINA, Badr et al. **A comparative study of decision tree ID3 and C4.5**. 2014. Disponível em:<http://saiconference.com/Downloads/SpecialIssueNo10/Paper_3-

KANTARDZIC, MEHMED. **Data Mining: Concepts, Models, Methods, and Algorithms**, 2nd. Canada: IEEE PRESS, 2011.

KORTING, Thales Sehn. **C4.5 algorithm and Multivariate Decision Trees**. 2014. Disponível em:<https://www.researchgate.net/publication/267945462_C45_algorithm_and_Multivariate_Decision_Trees>. Acesso em: 25 set. 2021.

LAROSE, Daniel T. **Discovering Knowledge in Data**: An Introduction to DATA MINING. Nova Jersey: John Wiley & Sons, Inc, 2005. 239 p.
LUGER, GEORGE F.; STUBBLEFIELD, WILLIAN A. **Artificial Intelligence (2nd ed): Structures and Strategies for Complex Problem Solving.** Pearson Education, 1993.

MAMAT, Nor Hana., et al. 2020. **Artificial neural network model and fuzzy logic control of dissolved oxygen in a bioreactor**. Indonesian Journal of Electrical Engineering and Computer Science.Vol. 17, No. 3, março 2020.

MITCHELL, Tom Michael. **Machine Learning**. 1. ed. Nova Iorque: McGraw-Hill Science/Engineering/Math, 1997. 432 p.

PATIL, Nilima; LATHI, Rekha; CHITRE, Vidya. **Comparison of C5.0 & CART Classification algotihms using pruning technique**. 2012. Disponível em:<https://www.researchgate.net/publication/284082342_Comparison_of_C5_0_CART_classification_algorithms_using_pruning_technique>. Acesso em: 10 out. 2021.

PAN, A., et al. 2020. **Association of Public Health Interventions with the Epidemiology of the COVID-19 Outbreak in Wuhan, China**. *Journal of the American Medical Association*,323(19), 1915-1923.doi:10.1001/jama.2020.6130.

QUINLAN, John Ross. **Improved Use of Continuous Attributes in C4.5.** Journal of Artificial Intelligence Research, 1996. Disponível em:<https://www.researchgate.net/publication/220543019_Improved_Use_of_Continuous_Attributes_in_C45>. Acesso em: 9 set. 2021.

ROKACH, Lior; MAIMON, Oded. **DATA MINING WITH DECISION TREES**: Theory and Applications. 2. ed. Londres: World Scientific Publishing CO. Pte. Ltd, 2015. 328 p. RUSSEL, STUART J; NORVIG, PETER; traduzido por SIMILLE, REGINA CÉLIA.

SINGH, Sonia; GIRI, Manoj. **Comparative Study Id3, Cart And C4.5 Decision Tree Algorithm**: A Survey. 2014. Disponível em:<http://www.academia.edu/34100170/Comparative_Study_Id3_Cart_And_C4.5_Decision_Tree_Algorithm_A_Survey>. Acesso em: 28 set. 2021.

SINGH, Sonia; GUPTA, Priyanka. **COMPARTIVE STUDY ID3, CART AND C4.5 DECISION TREE ALGORITHM**: A SURVEY. 2014. Disponível em:<http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.685.4929&rep=rep1&type=pdf>. Acesso em: 28 set. 2021.

SHMUELI, Galit et al. **DATA MINING FOR BUSINESS ANALYTICS**: Concepts, Techniques, And Applications With Jmp Pro. 1. ed. Nova Jersey: John Wiley & Sons, Inc, 2017. 467 p

SNEHA, N.; GANGIL, Tarun. 2019. **Analysis of diabetes mellitus for early prediction using optimal features selection**. Jornal of Big Data. Fevereiro de 2020.

TAN, Pang-Ning; STEINBACH, Michael; KUMAR, Vipin; **Introdução ao DATAMINING Mineração de dados.** Rio de Janeiro: Editora Ciência Moderna Ltda., 2009.

ORGANIZAÇÃO MUNDIAL DA SAÚDE. 2020. **Q&A: INFLUENZA AND COVID-19 - similarities and diferences**. 2021. Disponível em:

<https://www.who.int/emergencies/diseases/novel-coronavirus-2019/question-and-answers-hub/q-a-detail/q-a-similarities-and-differences-covid-19-and-influenza>. Acesso em: 10 out. 2021.

ORGANIZAÇÃO MUNDIAL DA SAÚDE.2020. **Coronavirus disease 2019 (COVID-19) Situation Report-48**. Disponível em: <https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200308-sitrep-48-covid-19.pdf?sfvrsn=16f7ccef_4>. Acesso em: 5 set.2021.

ORGANIZAÇÃO MUNDIAL DA SAÚDE. 2020. **Coronavirus disease 2019 (COVID-19) Situation Report-52**. Disponível em: <https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200312-sitrep-52-covid-19.pdf?sfvrsn=e2bfc9c0_4>. Acesso em: 5 set 2021.

ORGANIZAÇÃO MUNDIAL DA SAÚDE. 2020. **Numbers at a glance**. Disponível em: <https://www.who.int/emergencies/diseases/novel-coronavirus-2019>. Acesso em: 26 out 2021.

WIGUNA, Wildan; RIANA, Dwiza. 2020. **DIAGNOSIS OF CORONAVIRUS DISEASE 2019 (COVID - 19) SURVEILLANCE USING C4.5 ALGORITHM**. Jurnal PILAR Nusa Mandiri. Vol.16, No. Março de 2020.

WITTEN, Ian H.; FRANK, Eibe. **Data Mining: Pratical Machine Learning Tools and Techniques.** San Francisco: Morgan Kaufmann, 2005.