UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
FACULDADE DE ODONTOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM ODONTOLOGIA
MESTRADO EM ODONTOLOGIA, ÁREA DE SAÚDE BUCAL COLETIVA

ROGÉRIO DE FREITAS MOREIRA

IMPACTOS ECONÔMICOS DA COVID-19 NO MERCADO ODONTOLÓGICO PRIVADO NO ESTADO DO RS: O DESENVOLVIMENTO E AVALIAÇÃO DE UM *APP LOW CODE* DE DATA SCIENCE NO KNIME

Porto Alegre
2023

ROGÉRIO DE FREITAS MOREIRA

IMPACTOS ECONÔMICOS DA COVID-19 NO MERCADO ODONTOLÓGICO PRIVADO NO ESTADO DO RS: O DESENVOLVIMENTO E AVALIAÇÃO DE UM *APP LOW CODE* DE DATA SCIENCE NO KNIME

Dissertação apresentada ao Programa de Pós-Graduação em Odontologia da UFRGS como requisito parcial para a obtenção do título de Mestre em Odontologia – Área de Saúde Bucal Coletiva.

Orientador: Prof. Dr. Fernando Neves Hugo

Porto Alegre

2023

CIP - Catalogação na Publicação

Moreira, Rogério de Freitas
IMPACTOS ECONÔMICOS DA COVID-19 NO MERCADO ODONTOLÓGICO PRIVADO NO ESTADO DO RS: O DESENVOLVIMENTO E AVALIAÇÃO DE UM APP LOW CODE DE DATA SCIENCE NO KNIME / Rogério de Freitas Moreira. -- 2023.
157 f.
Orientador: Fernando Neves Hugo.

Dissertação (Mestrado) -- Universidade Federal do Rio Grande do Sul, Faculdade de Odontologia, Programa de Pós-Graduação em Odontologia, Porto Alegre, BR-RS, 2023.

1. COVID-19. 2. Fatores Econômicos. 3. Consultórios Odontológicos, Setor Privado. 4. Algoritmos. 5. Aprendizado de Máquina. I. Hugo, Fernando Neves, orient. II. Título.

Elaborada pelo Sistema de Geração Automática de Ficha Catalográfica da UFRGS com os dados fornecidos pelo(a) autor(a).

Agradecimentos

A Deus, que pôs no meu caminho este e outros desafios, alguns muito maiores do que outros, porém ao alcance de minhas forças e capacidades; que me deu resiliência, beirando a tenacidade (e mesmo a obstinação...), tudo na medida para que eu superasse estes desafios.

Ao meu orientador, o Prof. Dr. Fernando Neves Hugo, que me aceitou sob sua orientação, mesmo com uma proposta de pesquisa muito diversa da maior parte de sua produção acadêmica; e que foi forte o suficiente, mesmo em meus momentos de maior fraqueza, para manter a orientação, apesar de todos os meus sinais em contrário, jamais me faltando no suporte e apoio nas situações acadêmicas enfrentadas durante o mestrado. E que me deu a autonomia necessária para, usando minhas competências e potencial, prosseguir na pesquisa para aprender a converter minhas ingênuas pretensões na realidade construída ao longo deste processo.

À Fabíola Dornelles Molina, minha mulher, que esteve comigo este tempo todo, em todos os (muitos) *ups and downs* que enfrentei, me dando seu melhor apoio de diferentes modos e intensidades ao longo desta (não tão fácil) trajetória.

À Faculdade de Odontologia da UFRGS, ao Programa de Pós-Graduação em Odontologia (PPGODO) e sua equipe administrativa, por terem me aceito e apoiado durante o mestrado, que foi ainda mais prolongado e desafiador do que o previsto (em que pese a pandemia de COVID-19 que ocorreu durante a maior parte dele). A todos os "profes" do PPG-ODO, em particular aos Profs. Drs. Fabrício Collares, Cassiano Rösing e Tiago Fiorini, pelos ensinamentos na Disciplina de Metodologia Científica. Espero ter aplicado aqui alguns deles.

À equipe da Biblioteca Malvina V. Rosa, especialmente à Francieli Muck, por todo o apoio e orientações em meus problemas de busca em bases, formatação e muitos outros mais.

À Profª. Drª. Camila Santos, que me aceitou para o Estágio em Docência, e também me convidou para ministrar aulas a turmas de graduação em sucessivos semestres, todo o meu carinho pelas preciosas orientações e convivência neste período.

Aos colegas e amig@s do mestrado que tanto me apoiaram, em particular à Isadora Garcia, por ter me ajudado a aprender a fazer apresentações melhores e mais impactantes.

À Profª Drª. Suzy Camey (UFRGS), que me aceitou como seu aluno na Disciplina de Introdução à Ciência de Dados para a Saúde, e também me forneceu valiosas orientações estatísticas sobre como lidar com *missing values*. E ao querido Tiago Andres Vaz, que ministrou as aulas nesta mesma Disciplina e me apresentou a lógica e uso das GUIs, o que depois me permitiu construir todo o aplicativo descrito neste relatório de pesquisa.

Aos Profs. e pós-graduandos do Instituto de Informática e da Escola de Administração da UFRGS e da UFPel que me ajudaram, por terem dado toda a atenção e apoio, orientação e

contatos, especialmente ao Prof. Dr. Dante Barone e ao Mauricio Guntzel, no Orange; e às Profas. Flávia Azambuja e Daniela Brauner, bem como ao Gabriel Dornelles, no Tableau.

Aos Profs. Drs. Sílvio Cazella (UFCSPA), Fernanda Farinelli (UnB), Sérgio Dias (PUC-Minas) e Anderson Ferrugem (UFPel) pela disponibilidade e abertura a esta pesquisa.

Ao Beto Caffaro, Paulo Jeremias e Bruno Villar pelas explicações no uso de alguns *nodes* do Knime. Sem seu apoio e orientações, nas fases mais iniciais do meu aprendizado, quando eu ainda "engatinhava" na programação das etapas do processamento dos dados, eu nem teria concluído o ETL dos dados, e a pesquisa em ML sequer teria começado de verdade.

Aos funcionários, membros da equipe e participantes da Comunidade do Knime, sem os quais os segmentos mais complexos do workflow da pesquisa jamais teriam sido escritos; sem a generosidade e atenção de cada um, eu não teria conseguido construir este *app*. Alguns, por modéstia, pediram que não os citasse. Para evitar injustiças, agradeço juntamente a todos.

Ao Prof. Dr. Michael Berthold, pelas orientações e pelo próprio Knime, sobre cujas bases construí todo o meu trabalho. Ao Prof. Dr. David Hand, por seu apoio com as Curvas ROC. Ao Prof. Daniele Tonini, por ter dedicado parte de seu tempo e atenção aos meus (na época, ainda) incipientes passos no Knime. E à Drª Rosaria Silipo, cujas obras e orientações iniciais me estimularam a encontrar o apoio que eu buscava para a construção do workflow.

À CAPES, pelo suporte e apoio institucional enquanto bolsista durante parte do mestrado. O presente trabalho foi realizado com o apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001.

À ABO-RS, na pessoa de seu presidente, o Prof. Dr. João Batista Burzlaff, por seu apoio na divulgação da survey para esta pesquisa.

Peço desculpas aos muitos (ou melhor, muitíssimos) aqui não mencionados, que me ajudaram em maior ou menor grau, pois se eu nominasse a todos (caso conseguisse), o trecho dos "Agradecimentos" seria maior até do que o corpo da dissertação.

Agradeço até aos que me disseram que "... eu não sabia aonde queria chegar com a minha pesquisa...", que "... não sabia nem o quê procurava...", aos que zombaram de mim porque "... nem sequer devia ter tentado começar a fazer algo que não entendesse...". Sim, pois estes me ajudaram, mesmo sem o saber, a insistir, e ainda mais firmemente, em meus propósitos de tentar ajudar outros pesquisadores a se prepararem para pesquisar futuras circunstâncias comparáveis, usando esta ferramenta de DS (ou outras similares). Me ajudaram a resgatar, em mim e para mim mesmo, a força necessária para concluir esta tarefa.

A epígrafe de minha Epifania {;-:|)

Vivia mi'a rotina de dentista
Sem ter maior pendenga em meu caminho.
Tampouco novidades tinha à vista
Que dessem novo alento ao meu bom vinho.
'Té o dia em que eu achei-me nesta pista
Não tive mais conforto no meu ninho.
Já tive meus problemas de Gestão;
Quis eu, pois, ajudar outros, então.

Mas veio ante a mim a pandemia,
Turbou tudo o que eu antes pretendera.
Perdido vi-me eu, em agonia,
Jamais sonhara tal me sucedera:
Perdi o *timing* pro que eu mais queria,
Ninguém mais disse: "A meta inda é certeira!..."
E tive, mais de vez, que reinventar
O "Quê" e o "Como" eu ia investigar.

No zero foi meu ponto de partida
Rascunhos, não dispunha por suporte.
Confesso, dura foi a minha lida
E a base que eu dispunha não foi forte.
Ao Mestrado onde brota nova vida,
De Dados, a Ciência foi meu Norte,
E iluminou-me esta epifania,
Pro que eu buscava, trouxe a sinergia.

Tentei Tableau e Orange, foi vão.
Faltou, quiçá, a mim mais perspicácia.
Do Knime, enfim, me veio a salvação,
Pois nele eu pude achar mais acurácia.
Eu aprendi ao medo a dizer: "Não!"
Fazer um *app*, agora eu tenho a audácia...
E apoios recebi de toda parte:
Formação, expertise, amor e arte.

Ao não-saber, rejeito a sucumbência;
Da Ignorância, nunca mais a bordo!
Em Nau que leva rumo à sapiência,
D'antiga vida agora eu acordo.
Me trouxe a Academia em tal premência:
"Mate um dragão por dia...", e eu não discordo.
Mil temas mais, me vou a pesquisar,
Se a tanto a Data Science me ajudar.

Resumo

Introdução: Nesta pesquisa foi construída uma aplicação de Ciência de Dados na plataforma Knime, para a investigação e análise dos impactos econômicos da COVID-19 durante os seus primeiros 16 meses da pandemia, em consultórios odontológicos privados do Estado brasileiro do Rio Grande do Sul (RS).
Objetivo: Desenvolver e avaliar um aplicativo na plataforma analítica Knime visando identificar e interpretar relações entre diferentes variáveis, presentes em diferentes DBs com dados da distribuição e severidade dos casos de COVID-19 no RS, para inferir se estas variáveis estão associadas com métricas específicas relacionadas ao impacto econômico da pandemia no mercado-alvo.
Métodos: Todo o *app* foi construído no Knime, e todas as etapas foram executadas nele. Foram calculadas as seguintes métricas da COVID-19: a) Incidência, para mensurar a disseminação; b) Taxa de Hospitalização, para avaliar casos severos; e c) Taxa de Letalidade, para quantificar casos extremos. Os 497 municípios do Estado foram clusterizados segundo estas métricas mensais. Também foi aplicada em consultórios privados uma *survey* sobre retração do mercado. O cruzamento destas fontes de dados permitiu a seleção, configuração, aplicação e comparação da performance de 5 diferentes algoritmos de regressão (linear; polinomial de graus 2, 3 e 4; e logística) e 5 de classificação (*k*-NN; SVM; Naïve Bayes; MLP e AutoML) para avaliar quais deles tiveram os melhores desempenhos na predição dos efeitos econômicos em função das métricas da COVID-19.
Resultados: A pesquisa apresenta o workflow da construção deste aplicativo, e sua aplicação para: a) clusterização dos dados gerais da COVID-19, que gerou apenas 3 a 4 subgrupos que maximizem as similaridades intragrupo e dissimilaridades intergrupos; e b) predição dos níveis de acerto nas predições das métricas dos consultórios como função das métricas da COVID-19. Mostrou-se o desenvolvimento de uma abordagem no Knime para um processo de Descoberta de Conhecimento em Bases de Dados (KDD) pelo cruzamento de dados de diferentes fontes (aparentemente não relacionadas). A comparação entre performances, para este dataset, indicou que nenhum algoritmo de regressão, e nem o *k*-NN ou o SVM alcançaram boas performances; Naïve Bayes teve previsões apenas medianas; porém os algoritmos MLP e AutoML tiveram excelentes níveis de acerto nas predições feitas, respectivamente de 98,18% e de 100%.
Conclusão: O aplicativo foi construído com sucesso, porém a escassez de respostas à *survey* comprometeu a validade externa dos resultados da pesquisa, a qual permitiria predições amplamente acertadas e verificáveis no mercado-alvo. Não foi possível identificar associações estatísticas inequívocas entre as métricas da COVID-19 e dos fluxos nos consultórios. Porém ficou preservada a validade interna, legitimando a contento o objetivo da construção de todo um *app* no Knime, flexível e facilmente adaptável a outras questões de pesquisa e a outras bases de dados, mesmo por usuários sem treinamento formal em linguagens de programação. Outra meta atingida pela clusterização foi a de constatar a existência de 3 a 4 subgrupos a cujos membros pudessem ser aplicadas políticas similares no enfrentamento dos efeitos pesquisados.

Palavras-chave: COVID-19; Fatores Econômicos; Consultórios Odontológicos, Setor Privado; Machine Learning.

Abstract

Introduction: In this research a Data Science application was built on the Knime platform, for the investigation and analysis of the economic impacts of COVID-19 in the first 16 months of the pandemic, in private dental practices in the Brazilian state of Rio Grande do Sul (RS).
Objective: To develop and evaluate an application on the Knime analytic platform aiming to identify and to interpret relationships between different variables, present in different DBs with data on the distribution and severity of COVID-19 cases in RS, to infer whether these variables are associated with specific metrics related to economic impact on the target market.
Methods: The entire app was built in Knime, and all steps were run on it. The following COVID-19 metrics were calculated: a) Incidence, to measure dissemination; b) Hospitalization Rate, to assess severe cases; and c) Lethality Rate, to quantify extreme cases. The 497 municipalities in the State were clustered according to these monthly metrics. A survey on market shrinkage was also applied to private dental practices in RS. The cross-linking of these data sources allowed the selection, configuration, application, and performances comparison of 5 different regression algorithms (linear; polynomial of degrees 2, 3, and 4; and logistic) and 5 classification algorithms (k-NN; SVM; Naïve Bayes; MLP, and AutoML) to assess which of them had the best performance in predicting the economic effects as a function of the COVID-19 metrics.
Results: The research presents the workflow for constructing this application, and its use for: a) clustering the overall COVID-19 data, which generated only 3 to 4 subgroups that maximize intra-group similarities and inter-group dissimilarities; and b) predicting the hit levels in the predictions of the practices' metrics as a function of the COVID-19 metrics. It was shown the development of an approach in Knime for a Knowledge Discovery in Databases (KDD) process by cross-linking data from different (apparently unrelated) sources. The performance comparison for this dataset indicated that no regression algorithm, neither k-NN nor SVM achieved good performances; Naïve Bayes had only average predictions; but the MLP and AutoML algorithms had excellent hit levels on the predictions made, 98.18% and 100%, respectively.
Conclusion: The application was successfully built, but the scarcity of survey responses has compromised the external validity of this research's results, which would allow for largely accurate and verifiable predictions in the target market. It was not possible to identify unequivocal statistical associations between COVID-19 metrics and dental offices' inflows. However, the internal validity was preserved, legitimating the goal of building an entire app in Knime, flexible and easily adaptable to other research questions and other databases, even by users with no formal training in programming languages. Another goal achieved by the clustering was to verify the existence of 3 to 4 subgroups to whose members similar policies could be applied to address the studied effects .

Keywords: COVID-19; Economic Factors; Dental Offices, Private; Machine Learning.

Lista de Figuras

Lista de Tabelas

Lista de Gráficos

Lista de Siglas e Abreviaturas

AI – Artificial Intelligence
ANN – Artificial Neural Networks
APP – Application
AUC – Area Under the Curve [ROC]
CD – Cirurgião-Dentista
CNS – Conselho Nacional de Saúde
COVID-19 – Infecção viral gerada pelo SARS-CoV-2
CS – Coeficiente de Silhueta
CSM – Coeficiente de Silhueta Médio
DBs – Databases
DEE-RS – Departamento de Economia e Estatística do RS
DM – Data Mining
DS – Data Science
EDA – Exploratory Data Analysis
EPAO – Empresa Prestadora de Atendimento Odontológico
ETL – Extraction, Transformation, Loading
FP – Falsos Positivos
FPR – False Positive Rate
GUI – Graphical User Interface
HI – Health Informatics
JS – JavaScript
KDD – Knowledge Discovery in Databases
KNIME – +Konstanz Information Miner
k-NN – k Nearest Neighbors
LOD – Level of Detail
MA – Moving Average
MC – Matriz de Confusão
ML – Machine Learning
MLP – Multi-Layer Perceptron
MSE – Mean Square Error
NB – Naïve Bayes
NN – Neural Networks
PNN – Probabilistic Neural Networks
POV – Point of View
RBF – Radial Basis Function
ROC – Receiver Operating Characteristic [Curve]
RT-PCR – Reverse Transcriptase – Polymerase Chain Reaction
SARS ou SRAG – Severe Acute Respiratory Syndrome; Síndrome Respiratória Aguda Grave
SARS-CoV-2 – vírus causador da SRAG
SE – Standard Error
SES-RS – Secretaria Estadual de Saúde do RS
STARE-HI – STAtement on Reporting of Evaluation studies in Health Informatics
SVM – Support Vector Machines
TI – Tecnologia da Informação
TP – True Positive; TPR – True Positive Rate
VP – Verdadeiros Positivos
WIMP – Acrônimo de “Windows”, “Icons”, “Menus” e “Pointing devices”

SUMÁRIO

1 INTRODUÇÃO

Esta pesquisa apresenta a construção de um modelo de aplicação de Ciência de Dados (DS) na plataforma analítica Knime (Seção 2.3), para a investigação e análise dos impactos econômicos da pandemia de COVID-19 em consultórios e clínicas odontológicas privadas do Estado brasileiro do Rio Grande do Sul (RS), comparando bases de dados (DBs) abertas que descrevem o panorama geral da COVID-19 com os resultados de uma *survey* feita (no Google Forms), com gestores destes estabelecimentos odontológicos privados. Investigou-se como os diferentes locais em que estes profissionais e estabelecimentos atuam foram afetados de maneira muito diversa por variados efeitos da pandemia. E o presente trabalho busca o desenvolvimento de uma ferramenta que permitisse discriminar subconjuntos (cujos membros sejam mais comparáveis entre si do que com o conjunto geral dos dados) nos quais haja maior similaridade entre estes efeitos e potenciais determinantes ou diferenciais que os afetem, com as respectivas consequências nos mercados locais em que exercem suas atividades.

A presente pesquisa foi desenhada com a intenção de facilitar a capacitação de pesquisadores e/ou viabilizar tecnicamente outras pesquisas que queiram se beneficiar de estruturas semelhantes, e tratar problemas natureza próxima comparável à dos aqui tratados.

1.1 APRESENTAÇÃO, CONTEXTUALIZAÇÃO E DELIMITAÇÃO DO TEMA

As sociedades atuais sofreram em grande escala os impactos da pandemia de COVID-19, que afetou de maneira bastante diversa seus diferentes atores sociais. A maior parte dos setores econômicos sofreram uma retração de seus mercados, e os respectivos atores perceberam e interpretaram, sob diferentes escalas, os efeitos econômicos, sociais e de uso de recursos (públicos e privados) desta importante crise sanitária.

O segmento privado da Odontologia, tal como outros serviços (não públicos) de atendimento em saúde também pode ter sofrido uma retração em seu mercado, possivelmente associada a um maior temor, por parte de seus pacientes, de contaminação a partir do contato com ambientes frequentados por outros possíveis portadores do SARS-CoV-2 (COTRIN *et al.*, 2020; PELOSO *et al.*, 2020). Este foi um período considerado como amplo o suficiente para detectar alterações nas métricas dos fluxos de pacientes e financeiros destes estabelecimentos de saúde, *viz*., nos 16 meses de março de 2020 a junho de 2021, e desejou-se investigar possíveis

associações entre variações nestas métricas e dados mensurados por variáveis presentes em bases abertas sobre a COVID-19 no RS.

A continuidade dos atendimentos privados em Odontologia está associada à frequência e fluxos de pacientes (e, consequentemente, de fluxos financeiros) nestes estabelecimentos. Portanto, parece ser relevante a identificação e quantificação dos efeitos econômicos que a pandemia de COVID-19 trouxe sobre estes fluxos. Foram identificadas diferentes variáveis para quantificar diversos aspectos tanto nas métricas usadas para estes consultórios quanto nas DBs abertas com dados da COVID-19.

Pretendeu-se desenvolver um aplicativo no Knime que permitisse predizer se (e em quanto) as variáveis selecionadas podem servir para antecipar variações futuras nos fluxos que determinam a lucratividade, e suas consequências sobre o equilíbrio dinâmico de consultórios.

1.2 PROBLEMAS DE PESQUISA

A pesquisa aqui proposta foi orientada pelas seguintes questões:

a) Como as ferramentas de Ciência de Dados (em particular, o Knime) podem ajudar a descrever relações – especialmente as não muito evidentes ao se analisar grandes massas de dados – entre diversas variáveis presentes nas Bases de dados (DBs) abertas sobre o panorama geral da pandemia de COVID-19 no RS e retrações no mercado odontológico privado do RS?
b) Como os CDs e gestores de consultórios e clínicas odontológicas do RS medem e interpretam variações nos custos, nos rendimentos e demais impactos econômicos da pandemia de COVID-19 em seus estabelecimentos?
c) O desenvolvimento de pesquisas usando ferramentas de DS que levem à Extração de Conhecimento em Bases de Dados está ao alcance de pesquisadores iniciantes, mesmo sem formação prévia em linguagens formais de programação?

1.3 HIPÓTESES DE PESQUISA

A pesquisa envolve o teste das seguintes hipóteses, frente aos dados coletados:

a) Hipótese nula (H_0): a ferramenta Knime não permite a identificação de associações entre as variáveis que quantificam o panorama geral da COVID-19 no RS e as variáveis usadas como métricas para a movimentação de pacientes em consultórios odontológicos privados do RS.
b) Hipótese alternativa (H_1, ou de trabalho ou de falseamento da hipótese nula): no caso de a H_0 ser falsa, o Knime permite a identificação espacial e/ou temporal dos efeitos da COVID-19 no RS e sua relação com as métricas adotadas para avaliar as retrações no mercado odontológico privado do RS.

1.4 JUSTIFICATIVAS

Parece ser não apenas justificável, mas também imprescindível que os CDs e demais profissionais atuantes nos consultórios e clínicas odontológicas privadas do RS mantenham-se no desempenho de suas atividades regulares na Odontologia – observadas as restrições e cuidados adicionais relativos à potencial disseminação da doença – mesmo face aos problemas supracitados (Seção 1.1), decorrentes da pandemia de COVID-19. Estes problemas têm impacto direto na redução do faturamento global de cada estabelecimento. Esta redução pode ser associada à diminuição na busca dos pacientes por atendimentos ditos eletivos (*i.e.*, os que não derivam de situações de urgência clínica). E também ao maior tempo necessário nos novos procedimentos de desinfecção e de troca de barreiras microbiológicas entre atendimentos sucessivos. E ainda ao custo adicional destas últimas. Porém, pode haver outros fatores que levem à redução no faturamento geral dos consultórios. Portanto, parece ser importante compreender, mapear, fazer uma descrição mais consistente, adotando métricas adequadas para uma caracterização dos fatores e variáveis passíveis de mensuração, para auxiliar estes profissionais e gestores a se adequarem à "nova realidade" pós-COVID-19, e mesmo frente a outras possíveis grandes retrações (similares ou não) que venham a ocorrer neste mercado. Assim, busca-se identificar como manter a maior parte dos fatores positivos que tornam a prática privada da Odontologia uma profissão que desempenha um papel ativo na melhoria das condições gerais de saúde de uma sociedade.

1.5 OBJETIVOS

A pesquisa aqui apresentada visou o desenvolvimento e sua subsequente avaliação de um aplicativo construído na ferramenta de DS Knime, para o mapeamento dos casos de COVID-19 no RS, e das alterações de mercado decorrentes da doença, conforme percebidas por cirurgiões-dentistas (CDs) e por gestores de consultórios ou clínicas odontológicas privadas (ou Entidades Prestadoras de Assistência Odontológica (EPAOs)), durante os primeiros 16 meses da pandemia de COVID-19 no RS, *i.e.*, de março de 2020 a junho de 2021.

1.5.1 Objetivo Geral

Desenvolver e avaliar um aplicativo de Data Science na plataforma analítica Knime, para identificar e interpretar relações entre diferentes variáveis, presentes em diferentes massas de dados que estejam relacionadas à distribuição e à severidade dos casos de COVID-19 no RS visando inferir se estas variáveis têm influência em métricas específicas no mercado odontológico privado do RS.

1.5.2 Atividades Específicas

Para atingir o objetivo geral acima, a pesquisa buscou realizar as seguintes atividades:

a) pesquisar dados sobre a COVID-19 no RS em DBs abertas e fazer a seleção das variáveis mais associadas ao tema de pesquisa;
b) aplicar a ferramenta Knime para identificar e mapear padrões e associações entre as variáveis selecionadas;
c) fazer uma pesquisa do tipo *survey*, encaminhando um questionário online sobre retrações percebidas por CDs e/ou gestores dos diferentes consultórios e clínicas odontológicas privadas do RS com registro ativo no CRO-RS;
d) agrupar os municípios do RS em função de características comuns de susceptibilidade à COVID-19, conforme mensuradas pelas variáveis selecionadas;
e) buscar e interpretar associações significativas entre as variáveis selecionadas da COVID-19 e as do mercado-alvo;
f) buscar fazer todas as atividades de ML usando apenas os recursos *Low-code* ou *No-code* do Knime, *i.e.*, sem recorrer-se à necessidade de treinamento formal do pesquisador em linguagens de programação que usem linhas escritas de código;
g) aplicar e comparar alguns algoritmos de ML no Knime, para descobrir quais deles melhor se aplicam às massas de dados trabalhadas, e quais têm melhores performances preditivas.

2 REVISÃO DA LITERATURA

A Síndrome Respiratória Aguda Grave (SRAG ou SARS), em sua variante causada pelo vírus SARS-CoV-2, foi inicialmente identificada como uma forma de pneumonia severa, uma até então "nova" forma de SARS. Rapidamente foi caracterizado um "novo" tipo de Coronavírus (então chamado de nCoV) como o seu agente causador, e a doença atualmente é denominada como COVID-19. Ela teve origem em Wuhan, China, e foi anunciada no final de dezembro de 2019 (LI *et al.*, 2020; ZHU *et al.*, 2020). Trata-se de mais um evento de infecções virais em larga escala, determinados por patógenos emergentes ou reemergentes, sendo vários deles disseminados por vetores animais como os aves, morcegos e suínos, e sua trajetória evolutiva tem atraído o interesse de pesquisadores (GAO, 2018; LANA *et al.*, 2020).

Esta doença apresenta muitas características clínicas semelhantes a outras formas graves de pneumonia viral, como revisado por Guan *et al.* (2020), e por Wu e McGoogan (2020). Poucos meses depois, esta doença já tomava proporções inesperadamente altas, e passou a ter um impacto global, tanto em termos de saúde quanto econômicos (BALDWIN; WEDER, 2020). Em 11 de março de 2021, a Organização Mundial de Saúde (OMS) classificou a propagação mundial do vírus como Emergência de Saúde Pública de Importância Internacional (ESPII), em termos de uma pandemia (WHO, 2020). As teorias mais reconhecidas sobre a etiologia e modos de disseminação desta doença incluem transmissões iniciais de animais para humanos e subsequente transmissão sustentada diretamente entre humanos (CHAN *et al.*, 2020; OLIVAL; WEEKLEY; DASZAK, 2015).

Quando da declaração da OMS do estado de pandemia, conforme dados do *dashboard* sobre o panorama geral da disseminação da doença, disponível no site *Our World in Data*, o número de casos confirmados de contaminação pelo vírus era de praticamente 127 mil pessoas (126.969), distribuídas em 114 países, com um total, à época, de 4.618 mortes. Dados atualizados até o final do período coberto por esta pesquisa (30 de junho de 2021), ou aproximadamente apenas um ano e quatro meses depois, atingiram marcas globais de 183,15 milhões de casos confirmados de COVID-19, incluindo 3,96 milhões de mortes causadas pela doença. Estes números representam um aumento de cerca de 1500 vezes no número total de novos casos registrados e de mais de 85.600 vezes no número de óbitos causados pela doença. No entanto, estes dados globais – que podem ser acompanhados em diferentes fontes de boletins epidemiológicos, com dados atualizados diariamente, como no site acima, nos *dashboards* da OMS, do Our World in Data; da Johns Hopkins University; etc. – provavelmente não retratam

a grande taxa de subnotificação (IHME, 2022; RICHTERICH, 2020) no número de casos novos, seja por resultados inconclusivos, pela simples falta de material para testes, e mesmo pelo grande número de pessoas não submetidas ao teste padrão para detecção usados para confirmar a contaminação por este vírus, conforme já amplamente revisados por BÖGER *et al*. (2020) e crescentemente revisados e por outros autores e em publicações governamentais dedicadas e diariamente atualizadas (*e.g.*, as do Center for Disease Control and Prevention (CDC), nos EUA ou as do Ministério da Saúde, no Brasil.

Pouco tempo após firmado o *status* de pandemia, a OMS lançou um conjunto de orientações e procedimentos para que os diferentes países pudessem estar mais preparados para enfrentar os variados tipos de desdobramentos, em diversos cenários de saúde pública, frente a esta nova ameaça viral e às suas graves consequências (WHO, 2020). Outros pesquisadores também abordaram estas possibilidades de enfrentamento à pandemia em nível nacional (ANDERSON *et al.*, 2020). Alguns pesquisadores em diferentes países e instituições passaram a abordar diversos dos aspectos e desdobramentos de uma crise desta magnitude em estabelecimentos de saúde de forma geral, e particularmente nos odontológicos (COTRIN *et al.*, 2020; MENG; HUA; BIAN, 2020; PELOSO *et al.*, 2020; PEREIRA *et al.*, 2020).

Após a deflagração desta situação até então sem precedentes, e com a grande disponibilidade de ferramentas de comunicação e pesquisa para uso potencial em um quadro de semelhante magnitude, começou a ser feito um grande número de pesquisas sobre os mais diversos aspectos desta doença, tanto internacionais, como inicialmente revisado por De Melo *et al.* (2020), quanto brasileiros, e com evidências regularmente atualizadas pela Oxford-Brazil EBM Alliance. Esta é uma situação que somente pôde ser alterada após a implantação de campanhas de vacinação em massa, efetivas contra as diferentes variantes do vírus – tal como apresentado pela WHO e pelo CDC, apesar das grandes disparidades nas disponibilidades e aplicação de doses entre os diferentes países e continentes – terem participado das significativas reduções nos níveis e consequências desta pandemia. Apesar dos significativos avanços observados em alguns países, em outros ainda foi experimentada grande limitação na disponibilidade de vacinas contra o vírus. Mesmo os sistemas de saúde maiores e melhor estabelecidos, com maior expertise em amplas campanhas de vacinação, naquela época ainda anteviam – o que foi confirmado por dados dos períodos subsequentes – muitos meses até que pudesse ser feita uma cobertura suficientemente grande das respectivas populações nacionais e que seus efeitos imunizadores levassem a importantes reduções nos números totais de óbitos e nas médias móveis (MA) (*e.g.*, as de 14 dias) dos números de novos casos (em 24h) devidos a

esta doença. E, ainda frente à inexistência de alternativas terapêuticas eficazes para o tratamento dos pacientes com COVID-19, porém somente medidas de suporte à vida, os governos de diferentes países têm adotado maneiras peculiares para conter este quadro, cada um à sua época e panorama da doença, e segundo sua soberania nacional, sendo que as chamadas medidas de "distanciamento social", do uso intenso das barreiras sanitárias (*e.g.*, o uso de máscaras e de procedimentos adicionais de higienização) e de *lockdown* (de serviços não essenciais) parecem ter apresentado maior efetividade para minimizar tanto a disseminação da doença quanto as internações hospitalares (que sobrecarregam sistemas de saúde nos diversos países) e os óbitos decorrentes da COVID-19 (AMER *et al.*, 2021; VENKATESWARAN; DAMANI, 2020).

Embora não tenha sido a primeira pandemia enfrentada pela humanidade nos séculos precedentes (SILVEIRA, 2005; UJVARI, 2003, 2011), tal como outras pandemias respiratórias, seja no Brasil (BRASIL, MINISTÉRIO DA SAÚDE, 2005; COSTA; MERCHAN-HAMANN, 2016) ou em outros países (GRAY; SANDERS, 2020; IYENGAR *et al.*, 2020; POTTER, 2001; ZAMBON, 2014), a COVID-19 foi e está sendo vivida e quantificada com o uso de ferramentas de registro e de apoio para a tomada de decisão cada vez mais evoluídas (QUISPE-JULI *et al.*, 2020; SCHWENDICKE; SAMEK; KROIS, 2020; ZIMMERLING; CHEN, 2021), trazendo novas e importantes possibilidades para que as decisões tanto governamentais quanto dos diferentes agentes privados possam impor rumos que levem a uma redução da magnitude destes efeitos e em sua duração, com correspondentes aprendizados de importantes lições e grandes reduções nos números de vidas perdidas ou dos que experimentam sequelas da doença, tanto em serviços públicos quanto privados de saúde.

A despeito de (ou, talvez, alavancados por) todos os múltiplos e graves efeitos (pessoais, sociais, humanitários, econômicos, institucionais, etc.) desta pandemia, foram registrados alguns avanços tecnológicos no enfrentamento de problemas de saúde em larga escala para os quais ainda não havia expertise ou protocolos de ação previamente estabelecidos, como os registrados por Chandra *et. al.* (2022) em países ditos "emergentes e em desenvolvimento". O monitoramento do distanciamento social entre pessoas (potenciais disseminadores da doença) a partir de seus sinais de telefones celulares foi implementado em diferentes locais, como em SP (SANTOS *et al.*, 2021) e Porto Alegre (SOARES *et al.*, 2020). Tecnologias digitais também já haviam sido construídas para a melhoria na qualidade das integrações entre as DBs para o monitoramento dos dados da COVID-19 (HINKEL, 2022) e, anteriormente à COVID-19, para facilitar o monitoramento, a disseminação de informações (MAI *et al.*, 2017) e o enfrentamento e condução clínica de outras doenças transmissíveis inter-humanos, *e.g.*, como a tuberculose

(CAZELLA; FEYH; BEN, 2014). Dentre outras, inovações como as acima auxiliaram a motivar a construção do projeto da pesquisa aqui apresentada, que visou possibilitar um suporte adicional (baseado em DS) para decisões gerenciais frente a situações de crise sanitária, como a determinada pela COVID-19 no mercado-alvo da pesquisa (ou em outros comparáveis).

Uma das áreas do setor de serviços que apresentou uma grande retração, tanto em função dos fechamentos determinados por decisão governamental dos serviços não essenciais quanto pela redução na busca espontânea por atendimentos, foi a dos estabelecimentos privados de saúde, como nas consultas eletivas em clínicas e consultórios odontológicos. Alguns destes serviços passaram a atender somente a urgências, enquanto outros continuaram só com os tratamentos em andamento. As diferentes medidas de gradual abertura dos negócios de pequeno e médio porte experimentaram sucessivos relaxamentos e novas restrições, na medida em que as reaberturas eram acompanhadas por subsequentes aumentos nos números de novos casos comprovados de contaminação. Nos serviços odontológicos privados, órgãos governamentais publicaram decretos e normativas, e associações de classe prescreveram diferentes conjuntos de orientações e considerações como foi, nos EUA, o caso da American Dental Association (ADA), em suas considerações éticas sobre a reabertura de consultórios odontológicos durante a pandemia de COVID-19 (ADA, 2020). A ADA também fez uma *survey*, em períodos quinzenais, com consultórios odontológicos privados dos EUA, para mapear os impactos que estes sofreram com a COVID-19. No entanto, a citada pesquisa teve uma grande e rápida perda de respondentes ao longo do ano de 2020, o que reduziu em muito a validade de seus achados tanto quanto as conclusões a que seus autores chegaram.

A contaminação pelo SARS-COV-2 através de interações diretas e indiretas entre humanos inclui a transmissão entre dentistas (e suas equipes) e pacientes (CABRERA-TASAYCO *et al.*, 2020; CHECCHI *et al.*, 2020; PEREIRA *et al.*, 2020). As reduções observadas: a) nos agendamentos espontâneos de consultas; b) pelas restrições de funcionamento de serviços não-essenciais; c) pelas recomendações para adiamento de procedimentos não urgentes ou eletivos (como as do CDC (2020) para os estabelecimentos odontológicos); e d) pelo aumento dos custos e da duração de cada consulta (em função dos materiais e procedimentos de desinfecção intra- e interpaciente); são fatores que trouxeram importantes repercussões econômicas sobre a realidade destes consultórios (PATEL, 2020).

Ahuja (2020a, 2020b) discute interessantes abordagens sobre os aumentos de custos na era pós-COVID-19 e as mudanças que deveriam ser pensadas para as adaptações necessárias para minimizar os impactos econômicos nos estabelecimentos odontológicos. Este autor

argumenta que, mesmo que historicamente a Odontologia tenha sido uma das profissões mais caras, e com o aparentemente inevitável aumento dos custos da hora clínica, algumas mudanças sempre são (e foram) factíveis na mentalidade, no posicionamento, nos modelos de gestão e nas formas para manter uma Odontologia saudável tanto para pacientes quanto para profissionais, e ao mesmo tempo mantendo-se digna (e mesmo lucrativa) para estes últimos.

Acreditando-se no importante papel que a Odontologia representa dentro do panorama global para a saúde humana (HUGO; KASSEBAUM; BERNABÉ, 2021), e no poder que o uso de DS pode conferir aos tomadores de decisões que possam levar a diferentes alterações nos variados sistemas de saúde, incluindo-se aqui os da oral pública (DA CUNHA *et al.*, 2021), esta pesquisa aborda a atenção à saúde prestada em estabelecimentos não ligados aos sistemas públicos. A ideia deste trabalho foi concebida voltada ao desenvolvimento de um modelo flexível e aplicável em diferentes pesquisas, com múltiplas bases de dados (DBs) e para a investigação de variados tipos de questões de pesquisa. Pretende-se contribuir para as melhores condições de atendimentos em saúde oral, mesmo face a instabilidades geradas por fatores externos ao meio, e que perturbem o equilíbrio dinâmico deste segmento da profissão.

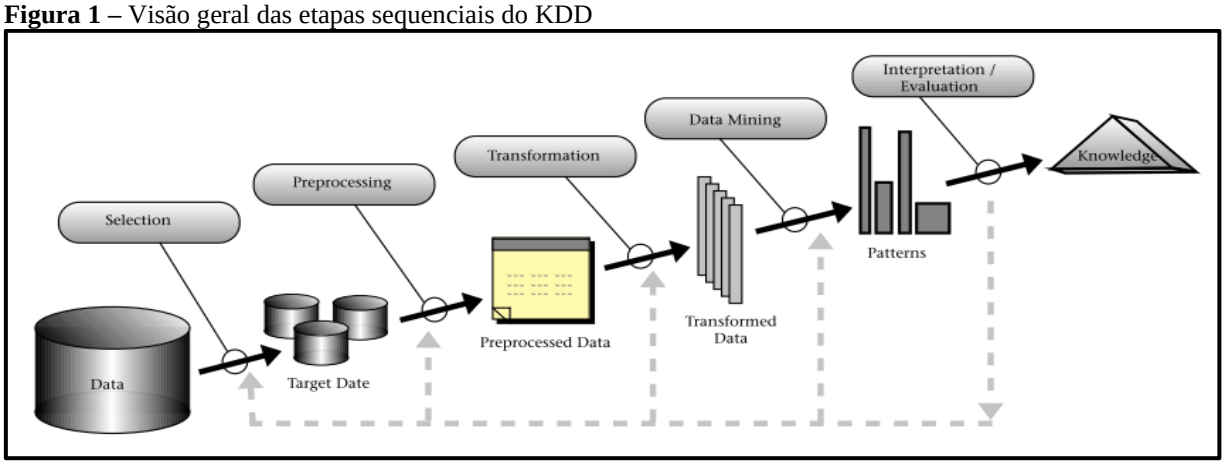
O presente trabalho visou a aplicação de uma ferramenta de Inteligência Artificial (AI) com interface gráfica (GUI) para a programação de variadas tarefas de Machine Learning (ML). A ferramenta escolhida (Seção 2.3) é a Plataforma Analítica Knime (BERTHOLD *et al.*, 2008). Nesta seção, será apresentada a justificativa para a escolha desta plataforma, em detrimento de possíveis outras.

Buscou-se a construção de um modelo que permita investigar as eventuais associações (não imediatamente evidentes) entre os dados – representados nas Bases de Dados (DBs) por meio de variáveis (categóricas ou numéricas) – ao analisar maiores massas de dados. A opção por uma abordagem de DS foi feita frente à grande massa de dados disponíveis sobre a COVID-19 em bases abertas. E o Knime foi selecionado, dentre várias outras ferramentas disponíveis, como o Orange Data Mining, o Tableau, o RStudio, e mesmo ferramentas mais tradicionais (embora também com mais limitações para tratar maiores massas de dados) como o MS-Excel ou o OpenOffice (GRECO, 2020), levando em conta a sua interface gráfica, que é bastante amigável e intuitiva, a diversidade de recursos oferecidos no Knime, sua maior capacidade e flexibilidade de processamento para a construção de programas mais complexos, e pela sua menor demanda de recursos, tanto locais de processamento e de memória local instalada (como no caso do Orange), quanto de conexão e largura de banda para processamento remoto (como no caso do Tableau). Outro ponto de maior atratividade para a seleção desta plataforma para o

desenvolvimento da presente pesquisa foi a maior disponibilidade de suporte oferecido pela supracitada “Comunidade”. Visou-se aqui um modelo flexível, para comparar atividades em saúde bastante diferentes entre si, e que possivelmente estejam inter-relacionadas (de modo uni- ou bilateral), e investigar as possíveis associações, avaliadas quantitativamente com dados extraídos de ambas as áreas. A adaptação deste modelo a outras pesquisas será favorecida, dentre os demais fatores já citados, devido à sua gratuidade. As ferramentas de DS tipicamente permitem investigações entre fenômenos mensurados por variáveis muito diversas, desde que atendidos alguns requisitos de vínculo entre as DBs que servirão como fontes de dados (MACÁRIO; BALDO, 2005). Comparações entre algumas das opções populares em artigos acadêmicos foram feitas por Ranjan, Agarwal e Venkatesan (2017) e por Venkateswarlu, Spanadna e Srikanth (2018).

Esta pesquisa visa permitir a descoberta de informações não imediatamente evidentes em grandes massas de dados, um processo que é geralmente referido como Descoberta de Conhecimento em Bases de Dados (KDD) (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996). O esquema das 5 etapas que compõem um processo de KDD é representado na Fig. 1.

Figura 1 – Visão geral das etapas sequenciais do KDD



Fonte: FAYYAD; PIATETSKY-SHAPIRO; SMYTH (1996).

Para as finalidades desta pesquisa, estas fases podem ser sucintamente descritas assim:

a) seleção ou segmentação – somente parte dos *raw data* (ou dados brutos, constantes das DBs originais) apresenta interesse para os objetivos do trabalho, e as demais não são incluídas;
b) pré-processamento – “limpeza” dos dados, remoção de inconsistências e *missing values* podem ser necessários antes de começar a trabalhar com os dados;
c) transformação – uniformização de formatos e padrão de todos os dados;
d) data mining – embora sendo descrita como a fase central, a mais importante de todo o processo, não pode ser implementada sem as anteriores terem sido adequadamente planejadas e executadas; corresponde à identificação e extração de padrões geralmente implícitos, não evidentes, em massas de dados que usualmente tenham maior contingente ou volume de variáveis e seus valores correspondentes; sendo normalmente automatizadas e executadas por

meio de algoritmos que efetuam a identificação dos padrões buscados, e que estão subjacentes aos dados a que estes algoritmos são aplicados;
e) interpretação e avaliação dos resultados – nesta fase é que os resultados anteriores podem ser sumarizados e que se considera o quanto de eficácia foi obtida nos processos executados, para julgar se o conhecimento a que se chegou, *i.e.*, qual o conhecimento que pôde ser extraído dos dados trabalhados e que pode ser considerado como verdadeiro.

A plataforma do Knime usa recursos de Inteligência Artificial (IA), na qual estão disponíveis diversas tarefas de Mineração de Dados (DM) e de Machine Learning (ML), conceitos estes amplamente revisados na obra clássica de Russel e Norvig (2013), mesmo muito antes do desenvolvimento de ferramentas gráficas para a construção de processos como os abordados e empregados ao longo da pesquisa aqui descrita.

Dois ou mais pesquisadores, frente à mesma questão de pesquisa, e mesmos recursos para o planejamento (ou programação) das atividades sequenciadas para atingir os objetivos fixados *a priori*, provavelmente idealizarão trajetórias conceituais e operacionais distintas. Estas podem mesmo atingir níveis de complexidade drasticamente diferentes. Mesmo quando ambas as soluções não pareçam diretamente comparáveis entre si, embora possam tomar os mesmos materiais de partida e chegar às mesmas soluções finais, os caminhos (e demanda de esforços) podem representar necessidades de investimento bastante diversos. Em uma dada investigação, o aumento na familiarização e na prática com os dados de partida e com os recursos disponíveis pode reduzir estes investimentos, levando a uma maior parcimônia no seu uso durante uma pesquisa, que tende a ser benéfica a todos. Onde possível, tentou-se aplicar aqui o "Princípio da Navalha de Ockham" (BORGES, 2022) nas diversas trilhas da programação. Em todos os segmentos do workflow para os quais foram identificadas duas ou mais sequências de operações que levem aos mesmos resultados, optou-se pela mais simples, mais curta, ou mais econômica (para o design ou para a execução destas operações).

Uma das simplificações para minimizar vieses originados por diferenças entre variadas sociedades, culturas ou hábitos de busca por atendimentos em saúde, foi o da restrição na inclusão de pesquisas feitas em sociedades marcadamente distintas da aqui pesquisada. Abaixo são citadas, à guisa de ilustração, algumas das escassas publicações localizadas (em periódicos acadêmicos), infelizmente não tendo sido possível identificar muitas publicações muito próximas do tema central desta pesquisa:

a) Garcés-Elías *et al.* (2022) pesquisaram os impactos (sociais, porém não econômicos) sobre dentistas das medidas de isolamento social como alternativa de enfrentamento à pandemia de COVID-19. Porém trabalharam com uma "amostra de conveniência" de uma *survey*, *i.e.*, não

probabilística. Embora ampla (*n* = 1195 CDs especialistas e clínicos gerais) e diversificada (com respondentes de 21 países da América Central e Caribe).
b) Cvetković *et al.* (2022) investigaram a construção de modelos, por meio de regressões multivariadas, para a identificação de fatores que possam servir como preditores para o temor relativo a COVID-19 (ou outras doenças pandêmicas), somente na República da Sérvia;
c) HARO *et al.* (2022) fizeram uma revisão da literatura sobre os impactos psicológicos que a COVID-19 teve nos ambientes públicos e privados da Odontologia em diferentes países, com 123 artigos (dentre os 3879 iniciais), dos quais 80 são relativos aos profissionais da Odontologia, e destes apenas 4 investigaram (exclusiva ou parcialmente) os aspectos econômicos das preocupações dos CDs decorrentes dos fatores econômicos diretamente ligados à pandemia e/ou da insegurança econômica;
d) BASTANI *et al.* (2021) fizeram uma revisão de escopo (em ago., 2020) incluindo 50 (dentre os 1553 iniciais) artigos sobre as preocupações globais destes profissionais de saúde oral, incluindo estratégias de continuidade no exercício de suas atividades, sendo que a primeira destas preocupações (em três estudos) foram as econômicas (devidas à perda de receitas e ao aumento nos custos de equipamentos de proteção e das rotinas de desinfecção).
e) CHAMORRO-PETRONACCI *et al.* (2020) aplicaram uma *survey* (abr., 2020) com 400 dentistas da Galicia (Espanha), durante o período de "Estado de Alarme", quando somente procedimentos de urgência eram autorizados, e estes profissionais acusaram fortes repercussões econômicas negativas quando mesmo máscaras FFP2 eram difíceis de obter.

Em resumo, a revisão de literatura feita não trouxe artigos muito próximos do tema desta pesquisa, nem no tema e nem na amplitude de tempo (para avaliações temporais), e tampouco incluindo simultaneamente o uso de ferramentas de DS.

2.1 ESPAÇO EUCLIDIANO MULTIDIMENSIONAL

Ao longo desta pesquisa, a Distância Euclidiana (ZHANG, 2008) foi usada em todos os *nodes* que avaliem distâncias entre pontos ou agrupamentos. A distância euclidiana *d* é provavelmente a maneira mais conhecida e empregada para avaliar distâncias em um espaço geométrico de *n* dimensões. Esta distância é tomada como um valor escalar, *i.e.*, seu módulo, sem sinal, e eis a razão de usar-se a raiz quadrada do quadrado das diferenças entre as coordenadas de cada ponto ao longo dos vetores das dimensões adotadas. Assim, elimina-se o risco de que distâncias opostas se anulem (ou se compensem), distorcendo o valor de sua soma. Assim, toda a metodologia empregada nesta pesquisa adotará esta concepção de que as grandezas mensuradas podem ser avaliadas ao longo de um espaço multidimensional, no qual cada variável representa uma dimensão deste espaço, e seus valores são operacionalizados vetorialmente. O principal motivo para a adoção da avaliação das distâncias entre pontos (ou observações) segundo diferentes vetores (*i.e.*, segundo um espaço vetorial) está na escolha metodológica do uso de um número arbitrável de variáveis, que foram definidas como as que

mais facilmente poderiam estar associadas às características epidemiológicas da doença (*i.e.*, as variáveis que poderiam descrever fatores determinantes (ou independentes)) e as que poderiam mensurar os efeitos da pandemia no mercado-alvo (*i.e.*, as variáveis que poderiam ser as dependentes).

A distância euclidiana *d* entre os pontos P = (p_1, p_2, ..., p_n) e Q = (q_1, q_2, ..., q_n), em um espaço euclidiano *n*-dimensional, é definida como:

a) $n = 1$ (espaço unidimensional, uma variável, um vetor linear). Para os pontos P = (p_x) e Q = (q_x), a distância *d* é calculada como:

$$d = \sqrt{(p_x - q_x)^2} = |p_x - q_x|$$

b) $n = 2$ (espaço bidimensional, duas variáveis, um plano cartesiano com pontos descritos por suas coordenadas em dois vetores perpendiculares). Para os pontos P = (p_x, p_y) e Q = (q_x, q_y), a distância *d* é calculada como:

$$d = \sqrt{(p_x - q_x)^2 + (p_y - q_y)^2}$$

c) $n = 3$ (espaço tridimensional, três variáveis, o espaço cartesiano convencional, com pontos descritos por suas coordenadas em três vetores perpendiculares). Para os pontos P = (p_x, p_y, p_z) e Q = (q_x, q_y, q_z), a distância *d* é calculada como:

$$d = \sqrt{(p_x - q_x)^2 + (p_y - q_y)^2 + (p_z - q_z)^2}$$

d) $n > 3$ (espaço *n*-dimensional, mais do que três variáveis, pontos descritos por suas coordenadas em mais do que três vetores, para além do espaço geométrico convencional). Para os pontos P = (p_x, p_y, ..., p_n) e Q = (q_x, q_y, ..., q_n), a distância *d* é calculada como:

$$d = \sqrt{(p_x - q_x)^2 + (p_y - q_y)^2 + \ ... + (p_n - q_n)^2}$$

Para a maioria dos problemas reais de pesquisa, uma ou mais variáveis centrais usadas para medir um determinado efeito sob estudo são dependentes de outras variáveis, e geralmente em número maior do que apenas 2 ou 3. Portanto, o espaço cartesiano convencional em duas ou três dimensões (2D ou 3D) geralmente não é suficiente para representar matematicamente as funções que descrevem as dependências que uma pesquisa identifica para as dependências de uma variável em função de 3 ou mais variáveis independentes. Em tais casos, é comum recorrer-se à abstração desta representação, *i.e.*, prescinde-se de uma representação visual destas funções, e recorre-se a simplificações para explicar os fenômenos em estudo, e consequentemente também as funções que os descrevem.

E, mesmo que as representações deste *n*-espaço (para $n > 3$) possam não coincidir com a consciência de experiência ordinária das pessoas, os resultados em maiores dimensões podem ser intuitiva e (quase) facilmente compreendidos e interpretados por meio de analogias semelhantes. Ou então recorrendo-se à substituição de variáveis, um processo pelo meio do

qual algumas destas dimensões são agrupadas em uma variável adicional (ou por uma das já presentes no *dataset* original), das quais ela seja dependente, e que possa substituir as variações das outras variáveis que ela represente. A esta variável que representa outras dá-se o nome de variável *proxy*[1]. Uma solução alternativa para a simplificação da análise de números mais elevados de dimensões está em um grupo de operações genericamente denominado de "redução da dimensionalidade" (BELLMAN, 1957), que pode ser feita por diferentes técnicas ou métodos. Esta redução de dimensionalidade pode envolver procedimentos de complexidade variável. Considerando-se que muitas dimensões são ou podem ser fortemente correlacionadas entre si, o uso de muitas dimensões (ou variáveis) pode ser desnecessário. E, por aumentar a necessidade de uso destes recursos ou técnicas, o excesso no número de dimensões também apresenta um maior custo (seja em termos de tempo e recursos de processamento, seja pelas maiores dificuldades práticas para a aferição dos valores), e ainda podem piorar o desempenho (*i.e.*, reduzir a acurácia) das estimativas e predições feitas. Cita-se aqui alguns destes processos:

a) um número arbitrário das variáveis originalmente presentes na DB de trabalho é agrupado em uma nova variável criada, cujo comportamento e variações represente o das demais, que é o procedimento de substituição de variáveis;
b) outro tipo de procedimento é aquele em que simplesmente se descarta o conjunto de variáveis que, embora mensurem aspectos importantes do fenômeno em estudo, representam variações não cruciais para as questões e objetivos da pesquisa;
c) um terceiro tipo é o da seleção de variáveis, que é o que foi adotado nesta pesquisa, no qual seleciona-se uma (ou poucas) das variáveis originais, e descarta-se todas as demais cujos valores tenham variações similares aos das que forem mantidas, e que possam ser representadas por elas, que são as chamadas "variáveis *proxy*".

2.1.1 Análise da independência das variáveis

A independência entre variáveis – o que é o mesmo que dizer que não há associação entre duas variáveis – foi avaliada entre cada par das citadas "taxas" da COVID-19 e entre cada uma destas e a métrica de fluxos nos consultórios. Tal como nos demais processos em que se adota o uso de uma variável *proxy*, em todos os algoritmos selecionados para a pesquisa aqui relatada (os que têm a independência de variáveis como um de seus pressupostos) esta independência foi identificada por meio de um dentre dois testes que expressam seus resultados respectivamente por meio de dois coeficientes:

[1] Uma variável "*proxy*" é aquela que pode substituir outra(s), para diferentes finalidades de avaliação, quando apresentar com estas variações similares em escala e direção, especialmente quando as demais forem de menor disponibilidade e/ou praticabilidade, ou ainda forem de maior dificuldade e/ou custo de mensuração

a) *r* de Pearson (1895), que investiga a ausência de associação linear entre as variáveis observadas. Este teste também tem como pressuposto a distribuição normal da amostra, que deve ser representada pelos valores absolutos das variáveis contínuas que a descrevem. Esta correlação foi investigada aplicando-se o *node Linear Correlation* do Knime.
b) ρ de Spearman (1904), que também investiga a ausência de associação entre as variáveis observadas, porém estas relações não necessariamente precisam ser lineares, e nem mesmo as variáveis precisam ser quantitativas, podendo, *e.g.*, ser variáveis ordinais. A correlação de Spearman, ao contrário do de Pearson, não tem pressuposições sobre a distribuição da amostra. Esta correlação foi investigada aplicando-se o *node Rank Correlation* do Knime.

Sob os limites de cada teste, estes coeficientes de correlação Pearson e de postos de Spearman (CALLEGARI-JACQUES, 2003) assumem valores no intervalo [-1; 1], para os quais "-1" indica uma correlação inversa e perfeita, "0" não indica correlação, e "1", indica uma correlação direta perfeita. E permitem a identificação de se existe uma associação linear (a) ou de outro tipo (b) entre pares de variáveis.

O design da presente pesquisa parte do pressuposto da investigação de se as três taxas selecionadas são (ou não) mutuamente independentes. Ou antes, pretende-se avaliar se há uma covariância (CALLEGARI-JACQUES, 2003) entre estas três taxas, *i.e.*, que tenham aumentos ou decréscimos alinhados entre si. Esta suposição será testada para os modelos por meio do teste de correlação entre as três taxas, através do uso do *node Linear Correlation* do Knime.

E, nas avaliações da independência entre as três taxas e a métrica dos consultórios, foi aplicada a avaliação pelo *node Rank Correlation*, adotando a configuração do *node* usando o coeficiente de correlação de postos ρ (ou *rho*) de Spearman, dados:

a) a escassez de respostas coletadas, o que levanta a suspeita de que as amostras não atendam aos requisitos da aleatoriedade e da representatividade (uma vez que para a maioria dos municípios, foi obtida apenas uma resposta, que tanto poderia ser a mais representativa da municipalidade quanto um outlier local);
b) que não se tem elementos que indiquem uma normalidade nesta distribuição; e
c) que a variável da métrica dos fluxos tem valores ordinais (*i.e.*, não contínuos, mas representando intervalos discretos).

2.2 TAREFAS PARA A IMPLEMENTAÇÃO DO MACHINE LEARNING

As sessões subsequentes são dedicadas a uma descrição e revisão das tarefas e etapas necessárias para a implementação de um processo de KDD, como é caso da presente pesquisa. Aqui são apresentados, de maneira mais sucinta, apenas os que foram nela adotados. E, dadas as suas características e configurações peculiares, o detalhamento de cada uma é apresentado no Cap.3, nas seções correspondentes a estas tarefas.

2.2.1 Procedimentos de Preparação dos Dados (ETL)

O pré-processamento dos dados selecionados destas DBs foram trabalhados no Knime por meio de processos de Extração, Transformação e Carregamento (ETL), visando uniformizar estes dados e selecioná-los ou modificá-los de acordo com os procedimentos alinhados com a busca determinada pelas questões de pesquisa.

Diferentes etapas sequenciais são seguidas ao longo deste processo, desde a importação de dados (de uma ou mais fontes de dados, sejam estas arquivos, DBs, sites, portais, etc.), sua seleção e pré-processamento, a transformação e uniformização dos dados para que sejam trabalhados dentro de uma ferramenta de DS (e a plataforma Knime contempla a realização de todas estas etapas), passando pela extração de padrões apresentados pelos valores das diferentes variáveis dentro dos ambientes estudados, até a análise e interpretação destes padrões, o que finalmente permite a extração do conhecimento buscado nas variações que descrevem os fenômenos-alvo de um projeto de KDD.

O desenvolvimento e a crescente aceitação da DS, tanto em ambientes acadêmicos quanto industriais ou comerciais, somente foi possível a partir do desenvolvimento e disseminação (a custos acessíveis) dos recursos técnicos necessários para operacionalizar o processamento e a análise de maiores massas de dados. E a DS parece ser particularmente adequada para integrar e analisar dados não estruturados[2] de diferentes fontes, para investigar e tratar problemas novos ou para os quais as soluções já conhecidas podem não ser suficientemente efetivas para gerar conhecimentos que levem à mudança de realidades concretas e complexas. Quando se dispõe de muitos dados, organizados segundo diferentes dimensões (*i.e.*, mensurados por diferentes variáveis), pode ser necessária a adoção de ferramentas com maiores recursos para processar os problemas enfrentados, traduzindo este(s) problema(s) para uma forma que permita a busca de soluções diretamente a partir das variações nestes diferentes dados e variáveis. Uma das tarefas que consomem mais tempo dentro de um projeto, em um aplicativo desenvolvido em uma ferramenta de Ciência de Dados, é a preparação destes dados, durante sua inserção nas ferramentas utilizadas para este fim. O conjunto de etapas do ETL visa extrair os dados das fontes a serem usadas, transformá-los e uniformizá-los em padrões nos quais estes dados possam ser operacionalizados no software, e finalmente carregá-

[2] Conforme sugere o nome, entende-se que estes dados não são apresentados conforme a mesma estrutura, seja em tipo de fonte, tamanho, formato, disposição gráfica, para os textos, ou mesmo tabelas e quadros para organizações tabulares de informações, ou ainda imagens (fotos, gráficos). As ferramentas de DS devem fazer sua importação e uniformização para que as informações desejadas possam ser extraídas deles durante o seu processamento.

los de maneira adequada ao caso, às questões da pesquisa e às DBs transformadas. As tarefas de ETL tipicamente chegam a consumir algo como mais de 80% do tempo total e do esforço empregados pelo pesquisador em um projeto de DS, de acordo com os dados (o processo dito como de *Data Mining* (DM)) carregados, com as questões a serem investigadas e com as técnicas de DS adotadas.

As ferramentas de Ciência de Dados geralmente têm melhor performance ao tratar dados numéricos. O modelo aqui construído apresenta a "rotulagem" de variáveis categóricas (com valores em *strings* (ou sequências de caracteres alfanuméricos)), substituindo-as por dígitos, de acordo com o Apêndice B. E, para as tarefas de ML, sempre que possível, as variáveis numéricas usadas são normalizadas – *i.e.*, transformadas para valores reais (do tipo *Double*) no intervalo [0; 1] – para converter todos os diferentes dados em valores em uma mesma escala, de forma a que valores absolutos muito diferentes não tenham um peso marcadamente diferente dos demais durante o processamento nos algoritmos, o que distorceria demasiado as possíveis interpretações dos resultados apresentados.

2.2.2 Análise Exploratória de Dados (EDA)

A Análise Exploratória de Dados (EDA) (TUKEY, 1977) é uma abordagem já clássica e bem estabelecida para que o pesquisador possa se familiarizar com os conjuntos de dados disponíveis e iniciar as análises subsequentes tendo em vista sua questão e objetivos de pesquisa. Quando se trata de um tema, aspecto ou abordagem ainda não bem definidos ou conhecidos, ou mesmo sobre os quais não se tenha certeza das relações entre as variáveis envolvidas nas DBs selecionadas, o pesquisador pode encontrar dados relevantes e prepará-los para análise. Porém, durante a realização da análise, estes dados podem não ser suficientes para a investigação das relações. E, assim, pode ser necessário recuar ou introduzir novas etapas e novos dados, descritos por outras variáveis, podendo então ser feito um novo mapeamento visual, o qual pode modificar substancialmente os insights até então obtidos no transcorrer da pesquisa.

Esta abordagem já contempla técnicas de Análise Visual, a qual pode ser descrita como uma ferramenta do raciocínio analítico que é facilitada pelo uso de interfaces interativas (como as adotadas no Knime), permitindo tratar mais adequadamente maiores tamanhos de massas de dados e que tenham relações complexas entre as variáveis, não evidentes em análises

superficiais. Seleciona-se algumas variáveis específicas (de uma ou mais diferentes DBs) e investiga-se as eventuais relações entre elas, por meio de ferramentas que permitam a identificação visual direta e interpretação de tais associações

Dada a importância desta fase inicial, a ferramenta escolhida para este mapeamento ou análise visual já contempla recursos que facilitem esta aproximação com os dados, conforme apresentado nas seções subsequentes, a começar por uma apresentação da ferramenta que foi usada, ao longo desta pesquisa, para a construção de todo o aplicativo desenvolvido.

Destaca-se também que, para os fins desta pesquisa, os termos "aplicativo" ("*app*", ou *application*), "solução em DS", são todos empregados como termos ou expressões equivalentes.

2.2.3 Processo de Clusterização

O processo de investigar massas de dados também pode ser referido como o de "fazer perguntas aos dados", *i.e.*, usar ferramentas de DS (*e.g.*, algoritmos de clusterização, regressão ou classificação) para identificar quais variáveis estão acompanhando mudanças em quais outras, ou quais as similaridades (ou dissimilaridades) entre variáveis dentre um número limitado de dimensões com valores observados. Cada "ponto" desta massa de dados é chamado de "observação", mas também pode ser uma pessoa, um paciente com uma patologia específica, ou uma dada patologia que acomete mais as pessoas com certas características ou comportamentos. Portanto, pode-se buscar associações em diferentes sentidos dos vetores que descrevem as observações. O processo pode ser executado com dados sobre os quais ainda não se disponha de suficientes informações para rotulá-los previamente à análise.

A clusterização (ou *clustering*) pode ser resumida como o processo de dividir um determinado *dataset* (ou conjunto de dados) em subgrupos com algumas características em comum, ou com valores de determinadas dimensões (*i.e.*, de variáveis) dentro de certos intervalos, que distingam cada subgrupo dos demais. A clusterização é um processo (ou tarefa) de *Machine Learning* (ML) (MADHULATHA, 2012) amplamente empregado em pesquisas que usem a DS.

Deseja-se obter as respostas mais simples possíveis para perguntas objetivas (que possam ser respondidas através das métricas selecionadas), como as de:

a) qual ponto é o mais representativo para um determinado subgrupo de pontos (*i.e.*, seu "centroide")?
b) quão coeso é cada subgrupo?

c) quais as distâncias entre os diferentes subgrupos e outros mais próximos em suas imediações?
d) qual o menor número possível de subgrupos (coesos entre si e distantes dos demais) que represente adequadamente esta massa de dados?

A tarefa de classificação de objetos é uma das mais primitivas da humanidade. Inclui desde a discriminação entre objetos, seres vivos ou locais que devem ser buscados (*e.g.*, para atender às necessidades de um indivíduo ou comunidade) e objetos, seres vivos ou locais que representam riscos indesejáveis. E esta distinção se propaga ao longo de toda a história humana, representando a essência de cada decisão cotidiana, em qualquer magnitude de importância, entre duas ou mais opções (mutuamente excludentes ou não). As pesquisas científicas, bem como as decisões gerenciais sobre quaisquer temas, não fogem a esta regra. Todo tomador de decisão (percebendo ou não o processo em que está incorrendo) necessariamente faz este tipo de discriminação a cada escolha feita, conscientemente ou não. Quanto maior o tamanho das massas de dados que baseiam uma decisão, mais simplificações devem ser feitas para se trazer do âmbito inconsciente para o racional e deliberado. Neste ponto da linha de argumentação é que se destaca a importância da clusterização.

Uma massa de dados *M*, que deva ser estudada para a resposta a qualquer questão de pesquisa (referente a estes dados), é tipicamente não-homogênea, ao menos de acordo com as escalas em que estes dados são avaliados pelas variáveis de interesse para estas questões, sejam estas variáveis contínuas ou categóricas. E, de acordo com os valores que os dados apresentem de acordo com estas escalas, a massa de dados pode, por finalidades práticas de análise, ser dividida em *k* subconjuntos (*subsets*), também chamados de *clusters*.

Um trecho importante da pesquisa aqui relatada é dedicado à busca por uma clusterização, em função da proposta de discriminar diferentes agrupamentos (ou subgrupos) na massa de dados, visando a extração de informações que subsidiem gestores para que possam tomar melhores decisões baseadas em resultados. A opção pela clusterização foi feita a partir da premissa de que medidas similares podem ser mais eficazes se aplicadas a grupos cujos elementos sejam mais similares entre si, o que pode torná-las mais efetivas, com uma melhor aplicação de investimentos ou de políticas específicas para estes subgrupos.

Recorda-se aqui uma máxima conhecida como “a primeira lei da Geografia” (TOBLER, 1970), segundo a qual "Todas as coisas estão relacionadas com todas as outras, mas coisas próximas estão mais relacionadas entre si do que coisas distantes". De acordo com esta concepção, parece fazer sentido analisar conjuntamente dados diferentes, porém de acordo com

suas similaridades. A adoção desta ideia está baseada no princípio de que uma maior similaridade entre os elementos de um grupo possa levar a uma maior efetividade nos resultados de medidas práticas aplicadas local e regionalmente. Entenda-se aqui que a proximidade pode ser indicada por suas “distâncias” ao longo de quaisquer que sejam as dimensões analisadas (sejam estas geográficas ou não), e ao longo dos vetores nos quais se deseja representar estas distâncias. Portanto, recorda-se que esta busca por similaridades segundo dadas dimensões, pode não ser em distâncias físicas, mas por outras variáveis (ver seções subsequentes).

As tarefas de ML permitem, por meio de recursos computacionais, o processamento de grandes massas de dados, analisando-os segundo múltiplas dimensões. Estes softwares (em particular, os chamados “algoritmos”) têm a capacidade de identificar e gerar visualizações e análises de padrões e associações não evidentes ante a multiplicidade de dados e dimensões disponíveis. Estes algoritmos podem ser genericamente divididos em dois grandes grupos, o dos “supervisionados” e o dos “não supervisionados”. Além destes, também há outras classificações (algumas delas híbridas) não abordadas aqui, como o dos “semi-supervisionados” e o dos “auto-supervisionados”. Nos supervisionados, o algoritmo “aprende", a partir de determinados critérios pré-estabelecidos, a discriminar quais dos elementos (observações ou “instâncias”) da massa de dados se ajustam melhor a um dos grupos simbolizados por elementos previamente já definidos como representantes destes grupos. Estes podem ser aplicados para classificação ou regressão. Os não supervisionados são geralmente empregados quando não se tem informações suficientes sobre os dados a serem analisados para “rotular” previamente alguns dos elementos já conhecidos desta massa (*i.e.*, identificar estes elementos como representativos de um subconjunto dos dados sob análise). Assim, o algoritmo testa iterativamente todos os dados, para identificar quais deles podem servir como “centroides” (ou centro representativo de um subgrupo de dados), geralmente a partir de um número pré-estabelecido de *clusters* a serem definidos. E, na etapa imediatamente subsequente, calcula (também iterativamente) as distâncias de cada ponto até cada um dos centroides, designando cada um destes pontos a um dos grupos já definidos.

O processo de clusterização é uma das tarefas de ML não supervisionadas que busca dividir uma massa de dados em subgrupos tão coesos (*i.e.*, cujos elementos sejam muito próximos entre si), e tão distantes uns dos outros quanto possível. Isto, claro, de acordo com as “distâncias” ou medidas ao longo das dimensões adotadas na análise. Novamente, recorda-se que cada uma destas dimensões é representada por uma das variáveis selecionadas. Assim, descreve-se aqui uma situação complexa de acordo com um número limitado de variáveis, na

busca de agrupamentos (*clusters*) segundo estas dimensões. Neste processo, geralmente se busca definir o menor número possível de *clusters* que se adequem a uma divisão suficientemente distinta de seus componentes em grupos cuja discriminação (segundo as variáveis analisadas) seja útil para a tomada de decisões baseadas nestes dados.

A Clusterização interpreta os dados de entrada, e busca grupos a que cada um dos dados possa ser associado a partir de sua semelhança com os demais elementos do grupo, segundo as dimensões escolhidas.

As aplicações mais específicas dos processos de *clustering* aqui adotados são descritas em maior detalhe nas seções correspondentes do capítulo 3, de Material e Métodos.

2.2.4 Algoritmos de Machine Learning

Para o desenvolvimento desta pesquisa, foram selecionados 10 (dez) diferentes algoritmos de ML, visando comparar-se as respectivas aplicabilidades às massas de dados trabalhados e as respectivas acurácias preditivas ou de classificação. Estes algoritmos foram 5 (cinco) algoritmos de regressão, *viz*., o de Regressão Linear, 3 (três) de Regressão Polinomial (em graus 2, 3 e 4), e o de Regressão Logística. E também foram escolhidos outros 5 (cinco) algoritmos de classificação, *viz*., o SVM, o *k*-NN, o Naïve Bayes, uma Rede Neural de Perceptron Multicamadas (desenvolvida especificamente para esta pesquisa) e o de AutoML (ou de *Machine Learning* Automatizada).

O uso, configurações e maior detalhamento destes algoritmos é apresentado nas seções correspondentes do Cap. 3, junto com sua aplicabilidade aos dados de trabalho e respectivos desempenhos (ou acurácias obtidas).

2.3 PLATAFORMA ANALÍTICA KNIME

O Knime (acrônimo de **K**o**N**stanz **I**nformation **M**in**E**r) é uma plataforma para análise de dados, aberta e colaborativa, desenvolvida na linguagem JavaScript (JS) e disponibilizada em *open source*, e que é gratuita (na maior parte de seus recursos). Este aprendizado foi muito facilitado pela ampla disponibilidade (na web) de tutoriais completos, e de diferentes guias (*e.g.*, FLANAGAN, 2011) e sites em que a sintaxe de scripts é didaticamente demonstrada mesmo para iniciantes no desenvolvimento. Esta disponibilidade decorre do uso muito amplo

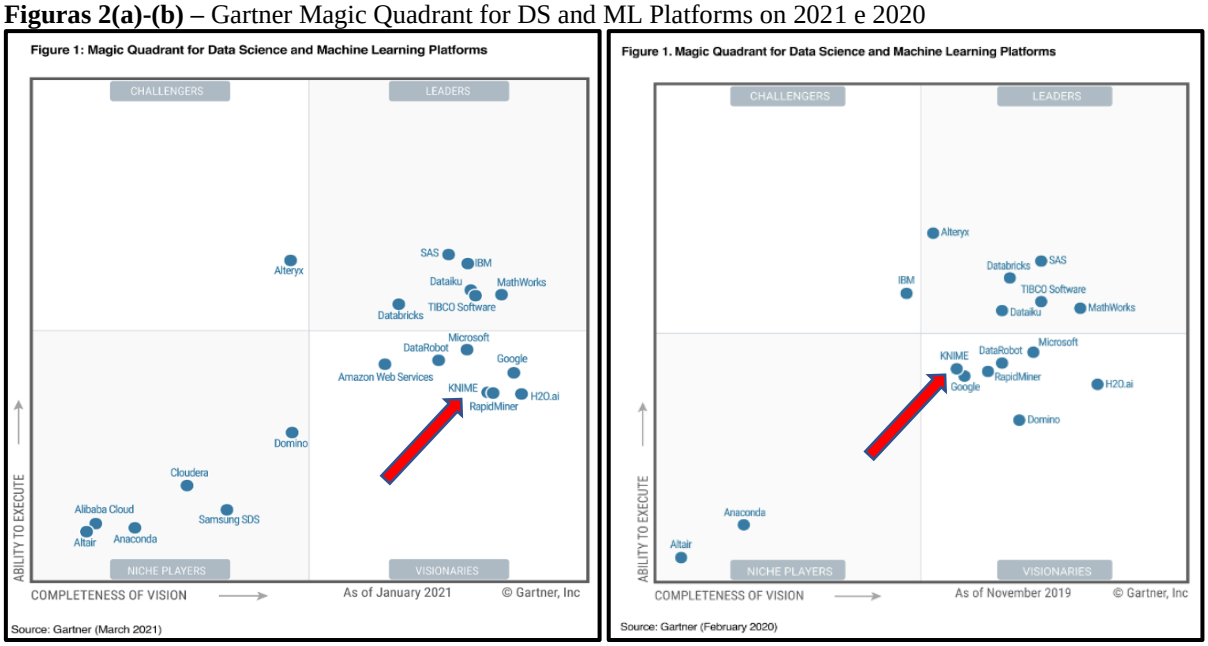
e bem estabelecido que tem esta linguagem. Os recursos do Knime são acessados através de uma interface gráfica, *i.e.*, por meio de *drag-and-drop* (ou "de arrastar e soltar"), através de ícones (cada um deles é chamado de "*node*", ou "nó") configuráveis diretamente pelos usuários, sendo que a maior parte das funcionalidades destes ícones são implementadas sem o uso de linhas de comando escritas, como nas linguagens formais de programação. Soluções de DS que permitem a programação por meio de interfaces gráficas, com pouco ou nenhum uso de linhas de código escrito são respectivamente chamadas de Ferramentas "*Low-code*" ou "*No-code*". Sucessivas evoluções e contribuições feitas ao Knime permitiram a integração com outras linguagens mais recentes como Python e R. No entanto, nos poucos casos esta pesquisa recorreu a linhas de código escrito (para a configuração de alguns *nodes*), isso foi feito em JavaScript.

As interfaces gráficas do usuário (GUI) demandam de seus usuários pouco ou mesmo nenhum conhecimento de linguagens formais de programação para tratar de diferentes problemas por abordagens computacionais. As interfaces gráficas, desenvolvidas e com aceitação progressivamente mais ampla durante as décadas de 1970 a 1990, foram inicialmente desenvolvidas para facilitar a interação Homem-Computador, flexibilizando (ou mesmo dispensando) o até então pré-requisito de digitação de texto para a inserção de linhas de comando para a execução de tarefas computacionais. O conceito básico das GUIs é representado pelo acrônimo WIMP, que indica uma interface composta por: a) *Windows* – uma (ou mais) janelas (independentes ou inter-relacionadas); b) *Icons* – ícones gráficos que representam as funcionalidades de cada elemento apresentado nas telas; c) *Menus* – das respectivas funções acessáveis ou executáveis em cada um destes ícones; e d) *Pointing devices* – acessórios ou elementos para a seleção e execução dos recursos disponíveis por meio gráfico. Uma das evoluções das GUIs, desde sua função inicial (a de ser dedicada apenas aos usuários de programas desenvolvidos por terceiros) foi a da possibilidade de construção de programas computacionais diretamente pelos usuários dos softwares que adotam GUIs. E o Knime oferece acesso à maior parte de seus recursos e funcionalidades em ícones gráficos (chamados de *nodes* (ou "nós")) para o processamento de dados através de poucos cliques para a configuração destes *nodes*, e (na maior parte de seus recursos) sem a necessidade de uso de linhas de código escrito em uma linguagem formal de programação. No entanto, sendo uma plataforma colaborativa, o Knime recebeu, ao longo de sua trajetória, a contribuição de muitos programadores independentes, o que agregou a possibilidade de uso, em vários de seus *nodes*, com a digitação de linhas de comando (*scripting*) em linguagens formais além do JavaScript, como Python e R (os *nodes* chamados "*snippets*"), ampliando significativamente suas funcionalidades.

As ferramentas de DS que adotam GUIs caracteristicamente são desenvolvidas contemplando elementos para esta familiarização inicial do pesquisador com os dados, *e.g.*, a EDA, trazendo elementos relacionados às subsequentes etapas e análises mais aprofundadas.

O Knime é de ampla aceitação e reconhecimento no ambiente de DS, a ponto de estar entre os integrantes do *Gartner Magic Quadrant for Data Science and Machine Learning Platforms* nos anos 2020 e 2021 (Figs. 2(a)-(b). Compete por mercado com Alteryx Designer, RapidMiner, SAS e o MatLab (e sua linguagem).

Figuras 2(a)-(b) – Gartner Magic Quadrant for DS and ML Platforms on 2021 e 2020



A plataforma Knime oferece diferentes cursos e ferramentas para treinamento e aprendizado (desde os seus níveis mais básicos), download gratuito de material instrucional (mesmo de cursos pagos), além de uma Comunidade ativa, um blog, um Hub (espaço virtual onde são postadas diferentes soluções e aplicações desenvolvidas para situações práticas de DM e ML no Knime). Também há um fórum no qual podem ser postadas dúvidas práticas, que é o espaço onde membros da empresa e outros usuários com maior expertise fazem contribuições para responder às dúvidas dos iniciantes ou menos experientes. Estas opções auxiliam muito na divulgação e disseminação da plataforma como sendo uma ferramenta útil para a solução de questões de pesquisa, para a análise e integração de dados, durante a investigação de problemas concretos, situações no mundo real.

Uma modalidade de trabalho recomendada – para um determinado problema de pesquisa, que envolva uma situação do mundo real e seja considerada como alvo de

investigação através do uso de Ciência de Dados – é a de se "fazer perguntas aos dados", *i.e.*, a de se manipular as variáveis mais pertinentes às questões de pesquisa, de forma a extrair, diretamente dos dados, associações que levem à obtenção de respostas para a indagação feita.

O Knime permite a importação de DBs em um variado número de formatos e fontes, e concede aos seus usuários uma ampla liberdade de manipulação das variáveis presentes nestes arquivos. Nele podem ser criadas DBs relacionais – *i.e.*, em que há uma ou mais variáveis (ou colunas) e linhas de dados em comum entre os diferentes arquivos, permitindo que dados inter-relacionados sejam buscados nos diversos arquivos a partir de quaisquer dos seus componentes – para a busca e identificação de relações e padrões subjacentes a estes dados.

O aprendizado inicial do Knime pode ser feito direta e gratuitamente, a partir de seu site de treinamento (e links de páginas relacionadas). Particularmente úteis na trajetória do autor da presente pesquisa para o aprendizado do uso dos recursos do Knime foram o Knime Hub e as contribuições individuais de alguns dos colaboradores no Forum Knime.

Todo o Knime é desenhado e operacionalizado por meio de WIMPs, distribuídos e interconectados formando um *workflow*, *i.e.*, um fluxograma, em uma área de trabalho (que, por *default* ocupa a maior parte da tela do Knime), e todas as (ou as principais) vantagens e recursos dos softwares que usam GUIs estão disponibilizadas por meio das guias laterais ou barras de menus superiores, e são operacionalizadas nesta tela.

Em aplicações mais extensas e/ou complexas, nas quais é implementada uma grande sequência de *nodes*, ou mesmo mais de uma sequência em paralelo, todo o workflow, se apresentado em uma tela, pode não ser imediata ou facilmente interpretável. Para tais situações, o Knime dispõe de um recurso adicional, que é o dos *Metanodes*, que são WIMPS que podem encapsular trechos do workflow com diferentes níveis de extensão e complexidade (ocultando-os em um único ícone) para facilitar uma visualização e compreensão mais simplificada das tarefas para os quais este workflow foi desenvolvido. Estes *metanodes* mantêm todas as funcionalidades dos trechos ocultos que representam.

Características similares aos *metanodes* são oferecidas pelo Knime nos *Components*. Tal como os *metanodes*, os *components* também são WIMPs que auxiliam na simplificação da apresentação e visualização de um workflow sem poluí-lo visualmente, porém com algumas vantagens adicionais, *e.g.*, sua possibilidade de configuração, o compartilhamento no servidor do Knime, e outras opções de visualização e contenção de *flow variables*.

Ambas as opções, *components* e *metanodes*, podem ser aninhados (*nested*), *i.e.*, estar incluídos em agrupamentos os quais, por sua vez, também podem ser aninhados em outros níveis de agrupamentos, cada um podendo conter outros *components* e/ou outros *metanodes*. Nos workflows construídos para a presente pesquisa, este tipo de aninhamento encadeado pode ser encontrado na Seção correspondente à construção dos workflows e apresentação dos resultados dos algoritmos de classificação. Este processo de *nesting* reduz em muito a sobrecarga visual, sempre podendo ser visualmente ampliado ou detalhado diretamente na área principal da tela da plataforma Knime.

Na pesquisa apresentada neste relatório, o workflow completo construído no Knime é apresentado, em toda a sua extensão e complexidade, no Apêndice C, embora cada um de seus trechos seja apresentado com maior nível de detalhe e tenha sido discutido junto ao texto referente a cada etapa ou tarefa de ML. Onde julgado adequado, foram incorporados *metanodes* e *components*, visando simplificar a visualização e compreensão destes trechos do workflow.

3 MATERIAL E MÉTODOS

Este capítulo descreve as considerações éticas e os procedimentos metodológicos adotados na presente pesquisa.

3.1 CONSIDERAÇÕES ÉTICAS

A participação na *survey* foi anônima, feita em uma plataforma online – a saber, a Google Forms. Isso garantiu, *a priori*, o total sigilo dos dados dos CDs e gestores participantes, e dos estabelecimentos em que atuam. A não identificação (ou anonimização) também atendeu o objetivo de descartar ou minimizar possíveis efeitos de constrangimentos, interferências ou vieses (associados à identificação de respondentes) na veracidade dos dados fornecidos. O envio dos convites de participação, feito diretamente pela equipe de pesquisa aos CDs e EPAOs, e também divulgado no site da ABO-RS, o que também serviu como uma garantia adicional de que as restrições impostas pela Lei Geral de Proteção de Dados (LGPD, Lei Nº 13.709, de 14 de agosto de 2018) (BRASIL, 2018) foram cumpridas com todo o zelo e rigor devidos. O anonimato das respostas buscou minimizar ou eliminar os riscos de vazamento de dados, que são típicos do ambiente virtual, e que estão além dos limites de atuação dos pesquisadores, sendo estes riscos gerenciados pela própria Google, que é uma instituição internacionalmente reconhecida quanto à funcionalidade e segurança de seus sistemas. Após a conclusão da pesquisa, os dados foram salvos em dispositivo local e apagados da plataforma.

Na primeira tela da *survey* online foi oferecido o Termo de Consentimento Livre e Esclarecido (TCLE) aos participantes da pesquisa, como parte integrante do próprio instrumento, que o acessaram após a leitura da "Carta-convite" (Apêndice A), para que os respondentes decidissem se: a) atendiam a todos os pré-requisitos para participação na pesquisa; b) sentiam-se livres para participar da pesquisa sem constrangimentos, sem interesses ou benefícios diretos oriundos desta participação. Ficou explícito que os participantes já teriam lido, compreendido, e estavam de acordo com o TCLE, uma vez que o acesso às telas do questionário foi habilitado somente após esta concordância. Recomendou-se que os participantes guardassem cópia dos documentos eletrônicos relacionados (Apêndice A; e Resolução 510 do CNS (BRASIL, 2012); (BRASIL, 2021)).

A pesquisa teve mínimo risco ou desconforto, e versou sobre tema já de interesse e questionamento regular de todos os profissionais da área, entre os quais os participantes desta pesquisa. A participação na *survey* implicou plena, imediata e total garantia aos respondentes de que, caso sentissem quaisquer riscos ou desconfortos durante a realização da pesquisa, a suspensão no preenchimento das questões antes do envio final das respostas assegurava a sua exclusão da amostra pesquisada, com a total e definitiva eliminação de todos os dados até então coletados.

O protocolo foi aprovado sob o CAAE: 47666021.2.0000.5347, pelo Comitê de Ética em Pesquisa da UFRGS (Sistema CEP/CONEP), um órgão colegiado, de caráter consultivo, deliberativo e educativo, cuja finalidade é avaliar, emitir pareceres e acompanhar os projetos de pesquisa envolvendo seres humanos realizados no âmbito da instituição, em aspectos éticos e metodológicos (Apêndice A).

A pesquisa teve como meta uma compreensão mais aprofundada do tema no cenário-alvo, foi de cunho estritamente acadêmico, usou abordagem quantitativa, e não teve finalidades comerciais. Os dados colhidos foram analisados e interpretados globalmente, servindo apenas para caracterização dos efeitos da pandemia de COVID-19 no mercado-alvo da pesquisa.

A elaboração da pesquisa visou os benefícios de fornecer ao ambiente acadêmico uma melhor compreensão: a) através da identificação dos efeitos da pandemia de COVID-19 observados no mercado odontológico privado do RS; b) quantitativa destes efeitos, com a meta potencial de subsidiar a elaboração de um preparo prévio para outras situações com natureza e dimensões comparáveis, para facilitar melhores condições futuras de permanência dos profissionais no exercício da Odontologia em seus estabelecimentos privados.

Os entrevistados declararam que participaram da pesquisa voluntariamente e sem remunerações de qualquer natureza e montante, que receberam apenas benefícios indiretos, por contribuir para os objetivos da pesquisa. Não houve custos de participação na pesquisa. Eventuais custos que os participantes tenham tido ao participar da pesquisa foram ressarcidos, desde que previamente combinado com os pesquisadores. A assinatura do TCLE não excluiu a possibilidade de que o participante buscasse indenização diante de eventuais danos decorrentes de participação na pesquisa, como preconiza a Resolução 466/12 do Conselho Nacional de Saúde (Conselho Nacional de Saúde, 2013).

3.2 POPULAÇÃO PESQUISADA NA *SURVEY*

A *survey* visou CDs e gestores odontológicos do mercado odontológico privado do RS.

Os CDs e (EPAOs) com dados de contato localizados receberam o convite de participação por e-mail, WhatsApp, uma rede social, ou outra via de comunicação eletrônica, conforme adaptado de Parashos, Morgan e Messer (2005).

Quanto aos critérios de inclusão, os potenciais participantes deveriam atender cumulativamente aos seguintes requisitos:

a) Atuar, exclusiva ou parcialmente, em Odontologia na iniciativa privada;
b) Exercer atividades de gestão no consultório ou clínica, seja na administração de todo o estabelecimento, seja na condução de carteira própria de pacientes, contabilidade, estoques, fluxo de caixa ou atividades relacionadas;
c) Os participantes não deveriam apresentar restrições ou conflitos de interesse com a presente pesquisa, que teve finalidade exclusivamente acadêmica e, portanto, sem vieses comerciais;
d) Ter seus rendimentos (no(s) estabelecimento(s) privado(s)) vinculados à produtividade ou rentabilidade do consultório, *i.e.*, não serem assalariados (com rendimentos mensais fixos).

A solicitação enviada à ABO-RS para a divulgação da pesquisa, e o convite para participação na pesquisa foi divulgado no site da ABO-RS.

Os profissionais foram diretamente contatados por diferentes meios eletrônicos e telefônicos, para o envio de cartas-convite de participação na pesquisa (mostradas no Apêndice A), contendo o link de acesso para participação na *survey*, por meio de:

a) aos consultórios e clínicas privadas que dispunham de um website, foram enviadas mensagens pelo canal do site;
b) aos consultórios e clínicas privadas que usam páginas de redes sociais, as cartas-convite foram enviadas pelo canal próprio da página;
c) parte dos profissionais foram contatados por e-mail;
d) parte dos profissionais receberam ligações telefônicas e os envios foram feitos para o canal indicado pelos profissionais.

O Instrumento de Pesquisa (Seção 3.4) foi desenvolvido usando-se o recurso de exibição ramificada das perguntas subsequentes com base nas respostas anteriores (que já é oferecido no Google Forms para o desenvolvimento de questionários), o que reduz o tempo necessário para que os participantes respondam à *survey*, e também diminui a desistência de participantes antes do seu preenchimento total e envio.

Este tipo de design de pesquisa (o da *survey*) permite que os participantes forneçam, em apenas uma ocasião, dados referentes a um longo período de tempo, como o que se deseja investigar, *i.e.*, os 16 primeiros meses da pandemia no RS.

3.3 BASES DE DADOS EMPREGADAS

Esta seção apresenta as DBs empregadas no presente estudo, sendo as quatro primeiras abertas (a)-(d) e as duas últimas fechadas (e):

a) o painel de dados da SES-RS (*.CSV) para o acompanhamento da situação da COVID-19 no RS;
b) um arquivo (*.XLSX) do Departamento de Economia e Estatística do RS (DEE-RS), com as populações estratificadas por faixa etária para os 497 municípios do RS (considerando-se que esta base é atualizada anualmente (nos meses de julho), assumiu-se, para as finalidades da presente pesquisa, como uma população constante, tomando-se o valor de 2020;
c) um arquivo (*.XLSX), manualmente compilado de diferentes páginas do site da SES-RS com a divisão dos municípios do RS em: 30 Regiões de Saúde; 21 Regiões COVID-RS; 18 Coordenadorias de Saúde - RS; e 7 Macrorregiões de saúde.
d) uma planilha do CRO-RS mostrando o número de profissionais (CDs e EPAOs) por município do RS;
e) duas planilhas do Google Sheets, que são produto da *survey*, respectivamente com a tabulação (pré- e pós-tratamento de dados, onde cabível, com a normalização das métricas) dos resultados das respostas aos questionários da *survey*, incluindo uma modificação destas com o acréscimo da divisão geopolítica descrita na alínea (d) acima. Estas duas DBs são as únicas a que se aplica o sigilo. As demais são abertas.

O reagrupamento citado acima, em (c), de algumas das Regiões de Saúde em Regiões COVID-RS foi regulamentado pelo Decretos Estaduais nº 55.240 e 55.428, respectivamente de 10 de maio de 2020 e de 06 de agosto de 2020, ambos disponíveis em página própria do Governo Estadual do RS como uma das medidas de enfrentamento à pandemia de COVID-19.

De cada uma destas DBs foram selecionadas algumas variáveis, com os seus subconjuntos específicos de valores usados para analisar as métricas escolhidas, e as consequentes “perguntas aos dados”, feitas por meio das ferramentas adotadas.

3.3.1 As variáveis selecionadas em cada DB

Para cada DB analisada, foram selecionadas algumas de suas variáveis originais, para a busca de relações entre elas e de associações estatísticas entre seus valores, conforme explicitado nas seções subsequentes.

3.3.2 DB dos dados gerais de COVID-19 no RS

Para a análise das DBs em 3.3(a)-(d) com os dados abertos da distribuição de COVID-19 no RS e das divisões geopolíticas e da administração da saúde pública em que o RS foi dividido pela SES-RS), foram escolhidas e renomeadas 9 das 30 colunas (ou variáveis) originais da primeira das DBs, conforme listadas no Quadro1, abaixo.

Quadro 1 – Variáveis selecionadas da DB sobre os casos de COVID-19 no RS

Variável	**Descrição sumária**
Municipality	Município de registro de cada caso
Reg_Covid	Reagrupamento da divisão político-administrativa da Saúde Pública do RS
Gender	Gênero do paciente
Age_Group	Faixa etária do paciente
Diagn_Criteria	Critério de diagnóstico usado para o registro do caso
Date_Confirm	Data de confirmação do caso
Hospitalized	Registro de se o paciente foi hospitalizado
Evolution	Registro da evolução do caso (Recuperação ou Óbito)
SARS	Registro do caso como sendo sintomático grave

3.3.3 DB com dados populacionais dos municípios gaúchos

Para a análise da DB em 3.3.c (um arquivo do Departamento de Economia e Estatística do RS (DEE-RS) com dados abertos sobre as populações dos municípios do RS, divididas por faixas etárias e por gênero. Para uma finalidade prática, a da comparação direta com os dados da DB com os dados gerais da COVID-19 no RS, as colunas desta DB foram recalculadas, exceto a de 0-4 anos (para a qual não foram obtidos dados separados entre os menores de 1 ano dos demais (de 1 a 4 anos)). Para as faixas etárias a partir dos 20 anos (que são apresentadas em faixas de 5 anos), elas foram somadas em pares subsequentes, com resultados em faixas decenais. As faixas resultantes, com total e subtotais por sexo, foram: de 0 a 4 anos; de 5 a 9 anos; de 10 a 14 anos; de 15 a 19 anos; de 20 a 29 anos; de 30 a 39 anos; de 40 a 49 anos; de

50 a 59 anos; de 60 a 69 anos; de 70 a 79 anos; 80 anos ou mais. Esta análise visou buscar subgrupos por gênero e/ou idade, para posteriormente usar esta DB para investigar se uma ou mais das variáveis que descrevem a distribuição da COVID-19 e de suas consequências no RS poderiam ser mais proeminentes em algum(ns) destes subgrupos.

3.3.4 Taxas adotadas para comparar dados e a normalização dos dados

O Quadro 2 mostra as taxas selecionadas para avaliar disseminação e severidade da pandemia no RS.

Quadro 2 – Taxas selecionadas avaliar disseminação e severidade da pandemia no RS

Taxa*	**Expectativa de justificativa para a inclusão da variável**
Incidência/10.000 Hab.**	Para a comparação de municípios com diferente n° de habitantes
Taxa de SRAG***	Taxa percentual da gravidade dos casos de COVID
Taxa de hospitalização****	Taxa percentual da gravidade dos casos de COVID
Taxa de óbitos*****	Taxa percentual da gravidade extrema dos casos de COVID

* foi trabalhada apenas na granularidade de "mês-ano" (embora originalmente disponível também com o dia);
** contagem relativa de novos casos na data (ou intervalo), em função da população local;
*** % dos casos com sintomas graves registrados;
**** % dos casos com hospitalização dos pacientes;
***** % dos casos com êxito fatal dos pacientes

Estas taxas foram usadas para tornar comparáveis os números absolutos de municípios com diferentes portes populacionais e de casos de COVID-19:

a) Incidência/10.000 hab. – a contagem (relativa a 10.000 habitantes no município) de novos casos em um dado período, o que torna comparáveis entre si a disseminação da doença em municípios com populações muito diversas;
b) Taxa de SRAG – o percentual dos casos diagnosticados com sintomas graves;
c) Taxa de Hospitalização – o percentual dos casos em que os pacientes são hospitalizados devido à gravidade dos sintomas;
d) Taxa de Letalidade – o percentual dos casos em que os pacientes falecem devido às consequências da doença.

Para poder tornar comparáveis os valores de diferentes variáveis que tenham valores absolutos muito díspares entre si, onde possível foi adotada a "normalização". A normalização é uma transformação afim muito frequentemente aplicada em estudos estatísticos, visando diminuir as discrepâncias durante comparações entre variáveis com valores absolutos muito distintos em escala (e.g., em uma comparação entre o PIB de um país e o salário de profissionais no mesmo país), na qual cada valor individual x_i, ao longo de uma dimensão específica, é convertido à distância correspondente em um dado intervalo. No caso da presente pesquisa, foi

adotado o método da "Normalização Min-Max" (também chamada de *Feature Scaling*), e foi escolhido como o intervalo fechado [0; 1] em todas as instâncias onde esta transformação foi aplicada. Este processo de transformação linear é feito mediante a seguinte fórmula:

$$x_t = \frac{x_i - x_{min}}{x_{max} - x_{min}}$$

onde:
x_t é o valor normalizado, *i.e.*, após a transformação pena Normalização Min-Max;
x_i é o valor original;
x_{max} é o maior valor absoluto original do conjunto a ser transformado; e
x_{min} é o menor valor absoluto original do conjunto a ser transformado.

Esta técnica de transformação linear converte todos os valores originais para os seus correspondentes neste intervalo. E simultaneamente preserva as relações de proporcionalidade entre os valores originais dos dados. O custo (ou contrapartida) da aplicação deste procedimento está em que ele também produzirá menores valores de desvio-padrão, o que também suprime o efeito de *outliers* sobre as demais medidas da dispersão em um dataset. Portanto, após a aplicação dos algoritmos (ou outros procedimentos) aos dados transformados, recorre-se a uma "desnormalização", o que devolve os valores às suas escalas originais. O Knime tem dois *nodes* específicos, um para cada uma destas funções: o Normalizer, que tem três opções de métodos de normalização: a) o "Min-Max", que já é o *default* do *node*, também com o intervalo *default* [0; 1], porém com a opção de configuração para outro intervalo arbitrário; b) o *z-score*, que realiza processo similar, mas convertendo cada valor original no número de desvios-padrão em que ele se afasta da média; e c) a normalização pela escala decimal, que divide cada valor original por uma potência de 10, até reduzir todos os valores à escala de trabalho desejada. Dos três métodos, a pesquisa adota o primeiro em todas as instâncias em que for aplicado. E o Knime também oferece o *node* Denormalizer, para a supracitada reconversão dos valores aos seus valores absolutos originais após a aplicação dos algoritmos pretendidos.

A decisão pela normalização dos valores foi tomada para minimizar a distorção ponderal que os algoritmos adotados atribuiriam às respostas que tratem de variáveis com valores absolutos muito distintos. Assim, pôde ser comparada a população do RS (da ordem de 11,5 milhão de habitantes) com o número de casos (cerca de 160 mil) em uma cidade como Porto Alegre – que tem cerca de 1.500.000 hab. distribuídos em cerca de 500 km² (*i.e.*, com uma densidade populacional de aproximadamente 3.000 hab/km²). Ou com outros dados totais, de magnitude muito diversa, evitando que os resultados sejam distorcidos em relação à

comparação entre a mesma população total do RS com a de qualquer outro de seus municípios menores – *e.g.*, Cacequi, que tem cerca de 12.500 hab. em uma área de quase 2.400 km² (com uma densidade populacional de pouco mais de 5 hab/km², ou 0,17% daquela da capital). Uma vez normalizados, os valores poderiam ser diretamente comparáveis usando os algoritmos selecionados, sem distorções causadas pela grande disparidade nos valores absolutos.

3.3.5 DB com dados das respostas da *survey*

As respostas aos questionários aplicados através da plataforma Google Forms sofreram um pré-tratamento, adaptando o conteúdo da planilha do Google Sheets que os continha, acrescentando as divisões geopolíticas em que o RS foi dividido pela SES-RS.

A pesquisa teve, como uma de suas metas, fazer um mapeamento mais efetivo sobre algumas características dos profissionais e empresas que atuam no segmento privado da Odontologia no RS. O Apêndice B apresenta as variáveis adotadas na *survey*, com os respectivos "rótulos"[3], geralmente no formato de *string*[4]. Estas variáveis envolvem:

a) características dos participantes (CDs/EPAOs): n° de profissionais, gênero, idade, tempo de credenciamento profissional; formação; carga horária e modo de atuação na Odontologia privada e em atividades de gestão;
b) características da clínica: localização; número de cadeiras odontológicas; equipe de CDs; especialidades atendidas; sistemas de pagamento por procedimentos;
c) período(s) em que o estabelecimento esteve fechado ou restrito a urgências;
d) variação percentual bimestral, no período investigado: no n° de atendimentos agendados e de procedimentos realizados; na fração de tempo ocupado em atendimentos clínicos; nos custos fixos e variáveis; no faturamento e lucratividade;
e) variação nas remunerações/retiradas financeiras de CDs, sejam estes sócios/proprietários, prestadores, locadores ou assalariados;
f) número de profissionais desligados da clínica;
g) outras medidas mais drásticas para enfrentamento da suposta retração de mercado em decorrência da pandemia de COVID-19;
h) mapeamento temporal e espacial da evolução epidemiológica da COVID-19 (por bandeira/região/mês), tanto em número de casos novos, hospitalizações e óbitos;
i) regiões em que o RS foi dividido, pela Secretaria Estadual de Saúde do Estado do RS, para a definição de políticas de enfrentamento à pandemia.

[3] Por "rotulagem", entende-se o processo de substituição de um valor (ou conjunto de valores), *e.g.*, a *string* de uma variável categórica, por outro valor, numérico (ainda que no formato de *string*), para simplificar o processo e minimizar os vieses gerados nos algoritmos pelo uso de *strings*).
[4] String é um tipo de dados em que o valor de uma variável é representado por sequências de caracteres alfanuméricos. Mesmo os algarismos não são passíveis de operações aritméticas ou algébricas, sendo apenas símbolos de outras grandezas.

Destaca-se aqui que este questionário teve questões para avaliar métricas das reduções nos fluxos de pacientes e financeiros dos consultórios privados, apresentadas em (d), acima. Estas métricas assumem apenas valores escalares, i.e., têm valores apenas positivos, pois são percentuais de redução nos fluxos. Assim, mensuram situações desfavoráveis, ou seja, mostram retrações de mercado, porém apresentam resultados com valores numéricos positivos. Quanto maior o valor, pior o efeito identificado. Este aspecto deve ser recordado na interpretação dos resultados ao longo das seções subsequentes, especialmente nos Caps. 4 e 5.

3.4 INSTRUMENTO DA *SURVEY*

O instrumento adotado na *survey* (CARRER *et al.*, 2020; MOHER *et al.*, 2014) foi um questionário com questões fechadas (tendo, ao final, uma questão aberta opcional, para dar espaço aos participantes para expor alguma opinião ou contribuição pessoal). Este questionário atendeu aos seguintes quesitos (FREITAS *et al.*, 2000):

a) as questões têm alternativas exaustivas, e ser mutuamente excludentes;
b) as questões são estritamente ligadas aos temas pesquisados;
c) a redação das questões deve ser pautada pelas implicações de seus enunciados nos procedimentos de tabulação e análise dos dados;
d) a redação das questões não deve gerar inconvenientes/constrangimentos aos respondentes;
e) a redação das questões deve ser clara, e alinhada com o nível cultural dos respondentes;
f) cada questão deve conter uma única ideia e possibilitar uma única interpretação;
g) o número de questões deve ser limitado, para não indispor ou cansar os respondentes;
h) deve iniciar com perguntas mais simples e sobre temas mais amplos, e deve progredir para as mais complexas e sobre tópicos mais focados; se possível, com encadeamento das ideias;
i) a redação de uma questão não deve induzir sua resposta nem a de outras questões próximas;
j) a apresentação gráfica do questionário deve buscar a facilitação de seu preenchimento;
k) os respondentes devem receber as devidas instruções de preenchimento, nomes e instituição dos pesquisadores, objetivos da pesquisa e importância de obter respostas corretas.

O questionário e o link para acesso a ele estiveram abertos e amplamente disponíveis aos interessados durante todo o período de coleta de dados da pesquisa.

O instrumento de pesquisa foi construído na e aplicado pela Google Forms, uma plataforma (online e gratuita) que pode ser usada para a realização de pesquisas de mercado. A eleição por esta opção também foi feita por outro fator atrativo, que está na credibilidade que um instrumento de pesquisa deste conglomerado de empresas (o *Alphabet, Inc.*) pode ter, frente aos potenciais participantes, na expectativa de aumentar suas taxas de aceitação e participação na pesquisa.

Durante a elaboração deste instrumento, foi adotada a ramificação das perguntas, para que os respondentes tivessem acesso direto às questões mais ligadas às respostas anteriores (e à sua própria realidade). Isto buscou facilitar a adesão dos participantes e o preenchimento das respostas e coleta de dados para o mapeamento dos atuantes neste mercado.

O questionário solicitava que os participantes classificassem as reduções observadas nos fluxos (de pacientes e financeiros) em um de seis intervalos percentuais, *viz*., < 5%; 5-10%; 11-24%; 25-50%; 51-75%; e >75%, respectivamente rotulados, para as finalidades de processamento nas ferramentas usadas nesta pesquisa, pelos valores alfanuméricos (do tipo *String*) 0,05; 0,1; 0,2; 0,4; 0,6; e 0,8 (Apêndice B).

3.5 ANÁLISE EXPLORATÓRIA DE DADOS DA COVID-19 NO KNIME

A EDA dos dados da COVID-19 foi feita no Knime através de uma série de etapas para a caracterização de alguns dos padrões de distribuição e severidade da COVID-19 no RS. Estas etapas estão encadeadas no trecho de workflow correspondente, conforme a Fig. 3.

Buscou-se um conhecimento mais aprofundado do pesquisador nos dados disponíveis, visando conhecer melhor o cenário geral da situação investigada, juntamente com a preparação do contexto conceitual sobre o qual foram executadas as etapas subsequentes, nas quais os dados foram descritos nas dimensões pretendidas para as análises feitas na pesquisa.

3.6 ALGORITMO DE CLUSTERIZAÇÃO *K-MEANS*

Uma das formas mais populares (*i.e.*, melhor conhecidas e mais usadas por pesquisadores) para a obtenção dos *clusters* é o algoritmo de *k-means* (MACQUEEN, 1967) – que deu origem a diversos outros algoritmos, após subsequentes modificações e melhorias – o qual divide massas de dados em *k* subgrupos chamados *clusters*, que são mutuamente excludentes e comparativamente diferentes entre si, cada um representado por seu centroide, nos quais se busca minimizar a dissimilaridade (ou maximizar a similaridade) *intracluster* (*i.e.*, identificar grupos que sejam mais coesos) e minimizar a similaridade *intercluster* (*i.e.*, identificar grupos que sejam mais distantes ou diferentes entre si), sendo ambos os critérios avaliados mediante as respectivas distâncias segundo as métricas adotadas. Estes *clusters*

podem auxiliar na explicação das características da distribuição de uma massa de dados e, assim, ser aplicados em diferentes técnicas de análise ou de mineração de dados. A aplicação padrão (*default*) do *k*-Means requer *a priori* uma definição arbitrária de um número de *subsets* em que se deseja dividir esta massa de dados.

Trata-se de um algoritmo de aprendizagem não supervisionada (N.; KAUTISH; PENG, 2021) – *i.e.*, em que os dados não são previamente rotulados, ou seja, que pouco se sabe de cada um deles além de suas coordenadas nas dimensões (ou métricas) em que estejam sendo representados no espaço analisado. Estes pontos (ou "observações") só poderão receber um "rótulo", *i.e.*, ser definido o seu "destino", ou grupo a que cada um dos pontos será alocado após ter sido feito o processo de classificação para o qual o algoritmo estará sendo "treinado". A expressão "treinamento de um algoritmo" refere-se ao processo iterativo no qual são ciclicamente testadas as possíveis combinações de variáveis e de distâncias até serem atendidos os critérios matemáticos previamente definidos. E estes ciclos se encerram após um número predefinido de execuções ou até que o algoritmo não produza mais alterações nas atribuições dos pontos a um dos *clusters* trabalhados, o que ocorrer antes.

3.6.1 Otimização no processo de clusterização

Como em qualquer outro processo de divisão de um grupo de dados em subgrupos, pode ser considerada a qualidade deste processo de reagrupamento, *i.e.*, a avaliação de quão similares entre si são os elementos de cada um dos subgrupos, e do quão diferentes cada subgrupo é dos demais. Uma dentre as diversas formas de avaliação desta qualidade é feita pela determinação do Coeficiente de Silhueta (CS) (ROUSSEEUW, 1987), que tem valores no intervalo fechado [-1, 1], onde valores mais altos, mais próximos de "1" indicam melhor ajuste entre os pontos e os *clusters* a que foram designados. Valores de CS mais próximos de "0" indicam uma neutralidade na atribuição de um dado ponto ao seu *cluster* apropriado, ou seja, os *clusters* estão muito sobrepostos. E valores de CS mais próximos de "-1" indicam atribuição ao *cluster* errado, *i.e.*, deveria estar agrupado em outro *cluster*. São calculados: a) o coeficiente de cada ponto, que indica quão ajustado o ponto é ao seu *cluster* ao mesmo tempo em que é distante dos *clusters* mais próximos; e b) o Coeficiente de Silhueta Médio (CSM) de cada *cluster*, que indica o quanto cada *cluster* é coeso e distinto dos de sua vizinhança. São computados dois valores, as distâncias médias *intracluster* e *intercluster*. O CSM foi a forma escolhida para esta avaliação de (dis)similaridade ao longo desta pesquisa e do seu processo de clusterização.

Embora a configuração *default* do *k*-Means preveja um número fixo para os *k clusters*, os CSMs representam uma das formas para achar o melhor número *k* para a clusterização, a aplicação iterativa de diferentes valores para *k* pôde ser feita (também no Knime, como se verá mais adiante), gerando *k* subgrupos que tenham os melhores CSMs possíveis.

3.6.2 Método do Cotovelo ("*Elbow Method*")

O processo de descoberta do melhor número, ou mais simplificadamente um número suficientemente adequado, dentre todos os possíveis *k* agrupamentos dos dados de um *dataset* para sua divisão em subgrupos de acordo com as dimensões analisadas e dentro do processo de KDD adotado pode envolver uma sequência relativamente complexa de etapas. É um conceito bastante empregado na Análise Visual, e o próprio nome remete a uma "curva abrupta ao longo de uma linha". Mais especificamente, adota-se a primeira curva mais abrupta ao longo da linha que representa a variação do CSM como função do número *k* de *clusters*. Muito frequentemente, este primeiro ponto de inflexão mais marcada é buscado como o primeiro "máximo local" (ou "mínimo local") (eventualmente, também o máximo (ou mínimo) global) desta curva. Esta primeira inflexão mais marcada representa um primeiro (e geralmente um suficientemente adequado) número de *clusters*, para o *clustering* (ou clusterização), dos pontos em grupos suficientemente coesos e distintos entre si, de forma que tenham maiores probabilidades de acerto as decisões tomadas com base nestes dados e, em particular, as decisões direcionadas a estes subgrupos específicos (KETCHEN JR.; SHOOK, 1996; THORNDIKE, 1953).

3.6.3 Atribuição (*assignment*) de pontos a cada *cluster*

Uma vez arbitrariamente definida a partição em *k clusters* (cada um representado por um centroide) para cada uma das "rodadas" do algoritmo, é feito o cálculo iterativo de quais foram os *k* pontos que podem ser estes centros, a partir das menores distâncias até cada um dos demais pontos.

Para a definição de quais pontos são alocados a quais *clusters*, calcula-se a distância entre cada ponto e outro ponto em sua proximidade. Para determinar qual ponto deve ser o centroide de seu agrupamento, calcula-se a distância *d* de modo similar ao descrito nas equações acima, considerando-se que o segundo elemento de cada par é o centro. E então é feita a

atribuição de cada ponto a somente um dos *k clusters*, de acordo com as menores distâncias que tenham aos respectivos centros. Cada ponto terá sua distância até o centroide do *cluster* mais próximo. Uma vez feita a alocação de cada ponto a um *cluster*, o processo é repetido sucessivas vezes, até ter-se atingida uma distribuição estável, *i.e.*, em que o aumento no número de rodadas não altere significativamente os pontos selecionados como *clusters*. Uma definição apurada de quais pontos são os mais adequados como centroides de seus grupos ocorre quando a distância *intracluster* (que os pontos são mais próximos do centro do grupo) é bastante maior do que a distância *intercluster* (o que indica que os grupos estão mais distantes entre si). Este processo de atribuição exclusiva de um ponto a somente um *cluster*. Dente outras possibilidades, esta foi a adotada para a pesquisa.

3.7 ALGORITMOS PREDITIVOS DE REGRESSÃO

Inicialmente, foi apresentado o trabalho com a massa de dados através do uso de técnicas de ML não-supervisionadas, *e.g.*, as de clusterização, para a obtenção de perfis de subgrupos dentro de um dataset maior, por meio de seu agrupamento de acordo com características em comum. Porém, as atividades de ML também incluem as tarefas ditas supervisionadas. Entre estas tarefas, aquelas em que já foi feita a definição dos grupos (com seus respectivos rótulos) aos quais os dados devem ser atribuídos, estão as de Predição. Ao contrário das não-supervisionadas, estas últimas têm como ponto de partida uma lista com um número finito de possíveis valores (ou intervalos) já conhecidos, e que são desenvolvidas para predizer valores a partir de um treinamento feito anteriormente. Elas podem ser genericamente divididas entre seus tipos principais:

a) de regressão – cujo resultado são valores numéricos atribuídos a novas observações descritas por variáveis que representam grandezas contínuas e tem padrões já identificados em dados similares, com os quais um algoritmo já tenha sido "treinado";
b) de classificação – as que permitem antecipar a classificação, *i.e.*, o enquadramento de um novo dado em classes ou grupos nos quais melhor se enquadrariam os valores de novas instâncias, também com base no treinamento feito, porém em uma classe (dentre um número finito de opções já conhecidas).

A maioria dos algoritmos supervisionados presta-se a um ou outro destes tipos de predição, embora já tenham sido desenvolvidos algoritmos configuráveis para a execução de ambos. No Knime, há mais tipos de algoritmos supervisionados do que não-supervisionados.

Foi feita a seleção e aplicação de 5 diferentes algoritmos de regressão e também de outros 5 de classificação, para a comparação das respectivas performances preditivas para o conjunto de dados coletados na *survey* para quantificar os efeitos da pandemia de COVID-19 sobre os consultórios odontológicos privados do RS, em função de três variáveis (as três taxas) usadas para mensurar a disseminação, a severidade e a severidade extrema da COVID-19, com o objetivo de investigar se, e neste caso, predizer em quanto, os fluxos (de pacientes e, por consequência, também financeiros) foram afetados pela pandemia neste período.

Para esta fase da pesquisa (a de ML), em todos os algoritmos de predição estes *datasets* são divididos – nos *nodes* *Partitioning* – em 70% para treinamento e 30% para predição. O objetivo desta parte da pesquisa está em comparar as métricas reais com as predições de cada algoritmo e quantificar quais destes últimos têm melhor performance.

Três destes algoritmos, *viz*., os de Regressão Linear, Polinomial e Logística, já são bastante conhecidos e empregados através de diferentes ferramentas, como os pacotes estatísticos (*e.g.*, Minitab, SPSS, etc.), e também têm versões disponíveis no Knime. Tal como seus nomes já antecipam, o primeiro visa descobrir associações próximas da colinearidade, enquanto o segundo busca a identificação de associações através de (ou descritas por equações de) polinômios em graus mais altos entre as variáveis analisadas (*e.g.*, relações de grau 2 (quadráticas), de grau 3 (cúbicas), etc.).

Tanto nas análises com o uso destes algoritmos de regressão quanto em quaisquer outras análises que interpretem níveis de significância de associações estatísticas, ao longo da presente pesquisa (exceto onde explicitamente indicado em contrário), quando usados, estes níveis de significância são representados pelo valor de sua probabilidade, sob a notação de seu *p*-valor[5]. Assim, quanto menores forem os *p*-valores em uma associação, maiores as probabilidades P (onde P = 1 - *p*) de a associação encontrada nos dados trabalhados ser representativa da população relativa a estes dados. Para os fins desta pesquisa, será adotado o *p*-valor padrão $p<0,05$ (*i.e.*, menor do que 5%) como indicando associações significativas. Este *p*-valor (<0,05) já é adotado como limiar (*threshold*) na maioria das pesquisas que avaliem significância das associações estatísticas entre dados de variáveis.

[5] O *p*-valor é um número real no intervalo fechado [0; 1], e corresponde à probabilidade de o resultado de uma análise, representado por este valor, ser devido ao acaso.

3.7.1 Predição dos valores usando o algoritmo da Regressão Linear

A análise dos dados pode identificar uma associação com coeficiente de correlação:

a) alto – *i.e.*, a associação é tanto maior quanto mais próximo o valor do coeficiente estiver dos extremos deste intervalo); e
b) significativo – *i.e.*, com *p*-valor (que representa a probabilidade de uma associação ser devida ao acaso, e tem valores no intervalo [0, 1]) abaixo de um limite pré-definido, entende-se que o conjunto dos dados pode ser adequadamente representado por uma reta. Para descobrir se há esta relação, aplica-se o algoritmo que busca traçar a melhor reta que passe entre estes pontos, e que os enquadre a uma distância relativamente pequena entre a reta e a maior parte dos pontos do *dataset*.

A reta assim traçada obedece a uma equação de primeiro grau (ou de primeira potência da variável (x) dita dependente, mostra-se como função de uma ou mais variáveis (y, w, ..., z) independentes dela. No caso de uma variável ser dependente de duas (ou mais) outras variáveis, trata-se de uma relação multivariada. Caso haja uma associação entre apenas uma variável independente e outra dependente, diz-se que há uma associação univariada. Os protótipos das equações que descrevem as relações lineares é:

$x = ay + bw + \ldots + h$; onde

x = variável dependente;
y; w; ...; z = variáveis independentes
a; b; …; d = coeficientes angulares das variáveis independentes; e
h = intercepto da curva no eixo das ordenadas

No Knime, este algoritmo é aplicado em um trecho do workflow que inicia com a partição dos dados no *node Partitioning*, e cada *subset* é conduzido para o *node* *Linear Regression Learner* ou para o *Regression Predictor*. Os dados de saída deste último são conectados aos *nodes* selecionados para a avaliação da performance do algoritmo, *i.e.*, que avaliarão os níveis de acerto e gerarão suas visualizações.

3.7.2 Predição dos valores usando algoritmos de Regressão Polinomial

Esta é uma investigação muito similar à acima, porém que considera que a curva que melhor descreve a resposta da variável dita dependente obedece a uma equação ou polinômio de grau igual ou superior a 2, como função de uma variável independente (para as associações

univariadas) ou como função de duas ou mais variáveis independentes (no caso das associações multivariadas).

Estas associações são identificadas quando a curva que mais se aproxima da maioria destes pontos seque uma equação de potência *P* mais elevada (com $P \geq 2$), pelo protótipo:

$y = ax^m + bx^n + cx + \ldots + pz^m + qz^n + rz + s,$

onde

y = variável dependente;
x; ...; *z* = variáveis independentes
m; ...; *n* = potências (ou graus) a que as variáveis são elevadas;
a; b; …; p; q; ... = coeficientes angulares das variáveis independentes; e
s = intercepto da curva no eixo das ordenadas

No Knime, este algoritmo é aplicado em um trecho do workflow que inicia com a partição dos dados no *node Partitioning*, e cada *subset* é conduzido para o *node* *Polynomial Regression Learner* ou para o *Regression Predictor*. Os dados de saída deste último são conectados aos *nodes* selecionados para a avaliação da performance do algoritmo, *i.e.*, que avaliarão os níveis de acerto e gerarão suas visualizações.

3.7.3 Predição dos valores usando o algoritmo de Regressão Logística

Este algoritmo realiza uma análise de Regressão Logística multinomial, definindo como coluna *target* a variável que apresenta os valores de resposta como função da(s) variável(eis) considerada(s) como independente(s). Para a aplicação desta regressão no Knime, são usados os *nodes* *Logistic Regression Learner* e *Logistic Regression Predictor*. Para isso, deve ser feita uma escolha (binária) de qual algoritmo (dentre duas opções) será usado para resolver o problema em análise. Na pesquisa será adotada a opção do *Iteratively Reweighted Least Squares* (ou dos Mínimos Quadrados Reponderados Iterativamente)

Na Regressão Logística (tal como em outros dos algoritmos aqui abordados), há alguns pressupostos que devem ser observados, como o da independência das variáveis, *i.e.*, que deve haver pouca ou nenhuma colinearidade entre variáveis independentes. Também deve ser considerado que este algoritmo tem melhor desempenho com maiores massas de dados.

3.8 ALGORITMOS PREDITIVOS DE CLASSIFICAÇÃO

Os algoritmos de regressão são geralmente mais adequados para massas de dados com variáveis quantitativas, *i.e.*, que medem grandezas contínuas. E, quando se pretende classificar (ou descobrir a qual classe ou grupo um dado ponto ou predição estaria mais adequadamente designado), como as variáveis categóricas, pode ser conveniente recorrer-se a algoritmos de classificação. Esta pesquisa testou algumas das opções disponíveis no Knime, como apresentado nas seções 3.8.3 a 3.8.5.

3.8.1 Procedimento para grandes desbalanceamentos entre classes em DBs

Para casos em que as DBs envolvem grandes desbalanceamentos entre as "classes" em que os dados (e subsequentes previsões) podem ser classificados, foi desenvolvido, no ambiente do Knime, o *node Synthetic Minority Over-sampling TEchnique* (SMOTE), onde são arbitrados dois parâmetros: a) os *k* vizinhos mais próximos aos dados que se quer prever, e são sinteticamente gerada combinações de valores próximos aos de suas ordenadas (nas dimensões em que estes pontos são descritos), e atribuídas as coordenadas dos novos pontos gerados (segundo as mesmas dimensões), permitindo que o modelo seja treinado com base em um maior número de pontos similares a cada ponto de treinamento; e b) o número de multiplicações (um número natural *n*, *i.e.*, uma entrada do tipo *Integer*) no subconjunto de treinamento (este número *n* é o "número de *over-sampling*"). Nas situações em que o SMOTE for empregado, é conveniente implementá-lo após a partição entre dados de treinamento e de teste, conectando-o somente às portas de entrada dos *nodes* LEARNER (e não às dos PREDICTOR (ou TEST)), para evitar aumentos artificiais nas previsões, sem os correspondentes aumentos em suas significâncias. Diferentes massas de dados podem ter melhores resultados usando *subsets* de treinamento aumentados em *n* vezes pelo SMOTE, porém sua aplicação deve ser testada previamente, e mantida somente nas situações em que a performance do treinamento for aumentada (*i.e.*, com redução nos erros de predição) pela sua adoção, por serem dados próximos aos reais, embora artificiais.

3.8.2 Procedimento para situações de *over-training*

Algumas massas de dados usadas em um projeto de ML são muito desbalanceadas. Outras, muito pequenas. Também podem ser usadas configurações – geralmente não intencionais, mas por falta de conhecimento sobre a situação geral que estes dados representam – em que um algoritmo aprende a prever com grande exatidão, mas apenas para os dados de treinamento, que são chamadas de situações de *over-fitting*, as quais mais frequentemente são o resultado de situações nas quais os parâmetros adotados nos *nodes* levam a um grande número de execuções do treinamento com os dados, porém gerando resultados bastante exatos apenas para estes dados de treino, mas não para novos dados com os quais o algoritmo ainda não tenha sido treinado. Em configurações deste tipo, é comum que os *nodes Learner* aprendam muito sobre um conjunto específico de dados, e as predições correspondentes sejam exageradamente precisas (*i.e.*, com acurácias muito altas) para aqueles dados específicos. No entanto, predições similares, porém feitas para novos dados (*i.e.*, outros dados que não os mesmos do treinamento, ainda que oriundas da mesma fonte ou DB) neste tipo de configuração mostram-se muito "pobres" (*i.e.*, com acurácias muito baixas). Uma das alternativas para averiguar se de fato ocorre um problema deste tipo é o da separação de uma pequena parte dos dados (tipicamente, algo como 10% do total) para a validação do treinamento, reservando o restante para a partição que já seria feita, entre os dados de treino e os de teste. Para os grupos de validação e de teste, o workflow é aplicado em paralelo, e a seguir é testado o grau de similaridade entre as predições no grupo de teste e no de validação. Em algoritmos cujas predições se assemelharam às situações de *over-training*, foi aplicada esta opção de validação, para avaliar se de fato estava ocorrendo um *over-training*, e, nestes casos, foi feita a redefinição dos intervalos adotados para os parâmetros dos algoritmos.

O Knime oferece o recurso de *loops* para a validação cruzada, através dos quais é feito um ciclo em que cada iteração é separada uma parte da massa de dados – usando o *node* *X-Partitioner* no início do loop e o *node* X-Aggregator ao seu final, sendo que este loop pode estar aninhado em outro loop maior, usado para o teste do desempenho sob diferentes parâmetros – e a execução é aplicada sobre esta massa, sendo os dados sequencialmente substituídos por outra parte dos dados. Usando-se esta abordagem iterativa, os ciclos são executados até ter sido percorrido o número (pré-definido) de iterações, e os valores são armazenados para a comparação subsequente. A finalidade deste ciclo é a de o treinamento poder ser verificado usando *n* diferentes partes dos dados e comparado com as predições. Deste modo, em alguns casos pode ser contornada a situação de *over-training*, pelo fato de o

treinamento poder ser executado com uma massa de dados composta pela somatória dos diferentes *subsets* de dados de cada iteração, também ciclicamente verificada para a somatória dos conjuntos de dados de validação, e a comparação entre as predições entre ambos apresentar uma melhor performance pelo uso em mais dados.

3.8.3 Algoritmo de classificação Support Vector Machine (SVM)

Um algoritmo bastante conhecido para a classificação de dados em subconjuntos é o das *Support Vector Machines* (SVM), originalmente chamado Support-Vector Networks (CORTES; VAPNIK, 1995), que foi desenvolvido para a busca da determinação de "hiperplanos" (com n - 1 dimensões) para a divisão de um espaço n-dimensional em dois subconjuntos, cada um contendo uma de duas classes em que os dados em análise foram agrupados. Assim, em um plano cartesiano tradicional (portanto, com $n = 2$), este hiperplano se reduziria a uma reta ($n = 1$); no espaço cartesiano convencional ($n = 3$), este hiperplano assume as características de um plano 2D); e, a partir de $n = 4$ (4; 5; ...; x), o hiperplano tem as n - 1 dimensões correspondentes (3; 4; ...; x-1), sempre com a separação entre dois subconjuntos. Conforme já argumentado, um dado problema pode ser descrito através de diferentes variáveis, cada uma representando uma dimensão. A resposta calculada pelo algoritmo é a do melhor plano (dentre os vários possíveis planos que transitem na região entre os subconjuntos), que dá apoio ("*support*") para a classificação feita, exatamente o vetor que maximize a distância entre os pontos mais próximos da região de vizinhança entre ambos. O algoritmo também tem como resposta outros possíveis hiperplanos, cada um segundo o vetor correspondente, e com as respectivas "margens" (ou distância entre este plano e os pontos extremos da periferia do *cluster*) (CRISTIANINI; SHAWE-TAYLOR, 2014).

O algoritmo SVM busca a melhor classificação dos dados entre dois grupos, e tem uma performance particularmente boa para problemas de classificação ou de regressão em que haja duas opções, ou duas classes. Porém, considera-se que é comum no mundo real o tipo de situação em que há problemas com mais do que duas classes – como é o caso da presente pesquisa, em que há seis diferentes classes. As perguntas do questionário (a *survey*) solicitavam que os participantes respondessem sobre as variações nas métricas de seus consultórios em 6 intervalos percentuais de redução nos fluxos, sob os "rótulos" que foram atribuídos a estas seis classes, *viz.*, respectivamente 0,05; 0,1; 0,2; 0,4; 0,6; e 0,8 (Apêndice B). Para situações como esta, com c classes (em que $c \geq 3$), o algoritmo SVM requer algumas modificações, para

iterativamente poder comparar uma das classes com as demais, alternando-se a variável, uma por iteração, até ser achada a combinação mais acertada entre os *nodes* e os trechos correspondentes no workflow.

O algoritmo poderia ser subsequentemente modificado para trabalhar simultaneamente com a classificação de dados em uma dentre mais do que duas classes, ou um *Multi-class* SVM (WESTON; WATKINS, 1998), computando o melhor hiperplano entre uma classe e todas as demais, porém esta opção acrescentaria um nível adicional de complexidade ao trecho correspondente do workflow, o que demandaria uma etapa adicional de operacionalização e outra de verificação da funcionalidade. Visando limitar estes aumentos na complexidade do workflow, pode-se adotar uma restrição adicional ao processamento durante esta etapa de ML, filtrando os dados para apenas duas classes, *e.g.*, aplicando uma solução à análoga à tabela gerada no processo de otimização do número *k* de *clusters* (que atribui o número do *cluster* a que cada ponto dos dados se ajusta melhor (*i.e.*, o de maior CS)).

O SVM requer, como entrada, uma das colunas com valores nominais, o que será trabalhado na seção de Resultados.

Os *nodes* do Knime para os SVM usam o algoritmo *Sequential Minimum Maximization* (SMO) (PLATT, 2000) e uma modificação feita nele (KEERTHI *et al.*, 2001).

3.8.4 Algoritmo de classificação *k*-NN

O *k Nearest Neighbors* (ou dos "*k* vizinhos mais próximos") é um algoritmo de ML supervisionada. Ele foi originalmente desenvolvido por Fix e Hodges (1951), e busca a classificação de um dado número de elementos baseado em características dos elementos de suas proximidades. Uma vez mais, ao usar o termo "proximidades" (ou "vizinhanças", ou "imediações") também aqui se recorre ao conceito de Distâncias Euclidianas, *i.e.*, calcula-se a diferença escalar entre as ordenadas de cada par de pontos de acordo com cada dimensão em que eles estejam representados. Tal como na já citada "1ª Lei da Geografia", este algoritmo supõe que deva haver mais semelhança entre pontos próximos do que entre pontos mais distantes. O *k*-NN é dito como um dos "algoritmos preguiçosos" (*Lazy Algorithms*) (AHA, 1997). Os *Lazy Learning Algorithms* exibem três características que os distinguem de outros algoritmos, que melhoram sua performance com o passar do tempo. Primeiramente, neles o processamento é retardado até que recebam novas solicitações de informação (*queries*); os *Lazy* simplesmente armazenam suas entradas de dados para uso futuro. A seguir, eles respondem à

query[6] combinando os dados já armazenados (*e.g.*, os de treinamento). E, finalmente, eles descartam as respostas construídas e quaisquer resultados intermediários. *Lazy algorithms* têm custo computacional mais baixo durante o treinamento, porém tipicamente têm maiores requisitos de armazenamento, além de custos computacionais maiores durante o processamento de respostas.

O principal parâmetro de entrada é representado pela variável *k*, ou o número de vizinhos mais próximos a serem considerados, por meio da qual o pesquisador define qual o número *k* de pontos das imediações de cada dado ponto que foram considerados quando se deseja classificar este ponto em uma classe (dentre duas ou mais). Para o modelo de classificação desenvolvido nesta pesquisa, recorda-se que são seis as classes em que cada dado de resposta (ou de saída) pode ser enquadrado. Ao contrário das "classificações *multi-label*", ou "multi-rótulo" (nas quais um dado ponto dos dados pode receber simultaneamente mais de uma classificação), na "classificação multi-classe" cada ponto pode receber somente um enquadramento, *i.e.*, as classes são mutuamente excludentes, porém há mais de duas classes para o enquadramento. O algoritmo, uma vez definido o parâmetro *k*, busca os *k* pontos mais próximos, e avalia qual o maior número destes pontos próximos que recebem a classificação na mesma classe. Quanto mais destes pontos próximos pertencerem a ela, maior será a probabilidade de que o ponto a ser classificado também pertença à mesma classe.

3.8.5 Algoritmo de classificação Naïve Bayes

O algoritmo de classificação multinomial Naïve Bayes (RUSSEL; NORVIG, 2013) está baseado no Teorema de Bayes, desenvolvido no séc. XVIII por Thomas Bayes, e que busca descobrir a probabilidade de que ocorra um dado evento (A) no futuro a partir do conhecimento da probabilidade já conhecida de ocorrência de outro evento (B) passado, desde que se saiba a probabilidade da ocorrência de A condicionada à probabilidade da ocorrência de B, segundo a equação:

$$P\ (A|B) = \frac{P\ (B|A).P\ (A)}{P\ (B)}$$, onde:

P (A|B) é a probabilidade condicional de A ocorrer, dado que se conhece a ocorrência de B;
P (A) é a probabilidade de A ocorrer;
P (B) é a probabilidade de B ocorrer;
P (B|A) é a probabilidade condicional de B ocorrer, dado que se conhece a ocorrência de A.

[6] Uma *query* é uma consulta ao banco de dados, a cada operação a ser feita.

O teorema (e a aplicação da técnica dele decorrente), podem ser adaptados para uma classificação multinomial, pois têm um pressuposto fundamental em comum, que é o da independência entre as variáveis tomadas como determinantes ou independentes. Em outras palavras, assume-se ingenuamente (daí o nome do algoritmo) que os determinantes não tenham correlação entre si, *i.e.*, que não se influenciem mutuamente.

Com a finalidade de evitar aumentos desnecessários (e/ou improdutivos, que não levem a uma maior discriminação nos resultados) na complexidade da elaboração do workflow do presente modelo, e em função das restrições implementadas no modelo para avaliar sua funcionalidade, a métrica *Effect(Mean*) (*i.e.*, a da variação percentual nas consultas efetivadas) será tomada como *proxy* das demais. Desta forma (caso exista uma violação do pressuposto da independência, após a verificação com o *node* correspondente), esta será contornada, recorrendo-se à substituição das três variáveis determinantes por apenas uma (dependente das demais), e assim o modelo terá que ser simplificado, para adequar a técnica original para a classificação pelo algoritmo.

3.8.6 Algoritmos das Redes Neurais

As Redes Neurais Artificiais (ANN) representam um estágio mais evolutivo da AI aplicada aos processos de classificação/predição, como os trabalhados ao longo desta pesquisa. As ANNs foram concebidas como uma tentativa de replicar (o que se sabe sobre) o funcionamento dos neurônios biológicos, que atuam recebendo informações – na forma de impulsos elétricos com origem determinada, identificada e em valor de intensidade reconhecida – processando-as e gerando resultados que, se superiores a um limiar (*threshold*) ou valor mínimo, devem ser conduzidos para o elemento seguinte da rede. Caso esta intensidade não ultrapasse o *threshold*, o sinal será considerado irrelevante.

As ANNs foram concebidas com muitas semelhanças ao funcionamento dos neurônios biológicos. Até mesmo as representações (gráficas) de neurônios, em uma ferramenta analítica que permita a construção de ANNs e que use uma GUI (como o Knime, entre outras), foram concebidas de maneira similar: apresentam conexões de entrada, em analogia aos dendritos de um neurônio biológico; um corpo central, onde as informações de entrada são processadas e as operações efetuadas pelo neurônio são desencadeadas; e uma ou mais conexões de saída, que

conduzem os resultados da operação de um neurônio artificial para o elemento seguinte da rede, em analogia aos axônios biológicos. Ao atravessar um neurônio, os sinais são atenuados ou amplificados, conforme critérios aprendidos (aqui chamados de "pesos") para aquela função que o neurônio executa, em analogia aos processos biológicos de memória. Os "pesos" são fatores pelos quais o valor de uma variável específica é multiplicado, visando a diminuição do valor do erro em uma predição ou classificação. Uma ANN é toda interconectada, e seus diferentes elementos (ou "neurônios artificiais") conectam-se a números específicos de neurônios em outros locais da rede, processam os respectivos sinais de entrada de maneira individual, aprendem limiares particulares para cada conjunto de valores com que trabalhem, utilizam diferentes pesos e realizam processamentos muito díspares em seus sinais de entrada, conforme a posição que ocupam na rede, e as funções para os quais foram desenvolvidos. Um "neurônio matemático" (ou "artificial") tenta reproduzir operações específicas, em analogia às do raciocínio (ou processo de tomada de decisão) biológico. Resumidamente, em termos de seus autores originais (MCCULLOCH; PITTS, 1943), eles recebem de maneira ponderada os sinais de entrada, aplicam a eles uma dada função, e os transmitem adiante para a próxima etapa da rede. Inicialmente, os autores acreditavam que um neurônio ou influenciava ou não influenciava o neurônio seguinte, de maneira binária ("*all-or-none*"). Posteriormente, após descobrir-se que não era assim para todos os casos, foram feitas sucessivas evoluções ou adaptações nas ANNs. Porém, a estrutura destes neurônios mantém o quesito de que a soma ponderada dos sinais de entrada ou dispara ou não dispara a função exercida pelo neurônio, se esta soma ultrapassar o limiar aprendido. E este "disparo" (*triggering*) ou não de uma função é realizado por uma "função de ativação".

As ANNs podem ser "treinadas" pelos processamentos dentro de cada neurônio e nos demais neurônios de uma rede, de forma que aumentos ou diminuições no valor do erro na operação de cada neurônio influencia (e de maneira ponderada, diferenciada) o processamento em cada um dos demais. Este "treinamento" corresponde a um ajuste iterativo dos pesos com que cada neurônio influencia seu próprio processamento e o dos neurônios das camadas adjacentes e interconectadas (IVAKHNENKO; LAPA, 1965). O treinamento original era iniciado no(s) neurônio(s) mais à esquerda (os de entrada) e propagava-se para a(s) camada(s) subsequentes até a dos neurônios de saída (os mais à direita no diagrama do workflow), em um modelo denominado "redes *feedforward*". Evoluções subsequentes foram feitas em 1970, originalmente por e a partir dos trabalhos de Linnainmaa (1976), com a adoção de uma retropropagação (*backpropagation*) das influências dos erros locais no ajuste dos pesos dos

neurônios das camadas adjacentes (uma antecessora e uma sucessora). Estes dois modelos conceituais ainda são os mais empregados.

O Knime disponibiliza duas opções de ANNs, tendo como base ambos os conceitos: a) as Redes Neurais Probabilísticas (PNN) (SPECHT, 1990), que seguem o conceito de *feedforward* e usam o algoritmo do Ajuste Dinâmico de Decaimento (DDA) (BERTHOLD; DIAMOND, 1994); e b) os Perceptrons Multicamadas (MLP), que segue o conceito de *backpropagation*, treinados usando o algoritmo RPROP (RIEDMILLER; BRAUN, 1993).

3.8.7 Algoritmos das Redes Neurais – as Redes Neurais Probabilísticas (PNN)

Nesta pesquisa não foi usada a segunda opção do Knime para Redes Neurais (NN), que é a das Redes Neurais Probabilísticas (PNN). Sendo tão limitado o *dataset* de trabalho nesta pesquisa, seria menos indicado trabalhar com uma NN que fosse apenas *feedforward*.

3.8.8 Algoritmos das Redes Neurais – os Perceptrons Multicamadas (MLP)

As ANNs usadas nesta pesquisa são os MLPs, que são construídas com:

a) uma camada de neurônios de entrada – geralmente, os modelos construídos segundo este conceito têm um neurônio para cada variável na camada de entrada. No caso apresentado aqui, estas variáveis são as três taxas da disseminação e gravidade da COVID-19;
b) uma ou mais camadas ocultas de neurônios (i.e., em que não é possível prever a saída desejada), e nas quais são iterativamente atribuídos pesos para cada um dos componentes de erro, visando minimizar a taxa de erro na classificação final;
c) uma camada de neurônios de saída, sendo que, nesta última, geralmente é descrito um neurônio para cada uma das classes de resposta já conhecidas, nas quais se deseja fazer a previsão do enquadramento de cada observação de entrada. Assim, são seis os neurônios de saída, cada um correspondendo a uma das seis classes ou faixas percentuais de variação nos fluxos de pacientes.

Os MLPs aplicam o algoritmo RPROP, em que há um ajuste de retropropagação (*backpropagation*) iterativa para a minimização dos Erros Quadráticos Médios (MSE). O MSE é um estimador de um preditor que: 1°) calcula as diferenças entre cada um dos valores preditos e o valor real correspondente (nos *subsets* de treinamento e de validação); 2°) soma estas parcelas; 3°) eleva este resultado ao quadrado; e 4°) divide este valor pelo número de elementos preditos. Quanto maior for este valor (*i.e.*, o do MSE), pior será o modelo, ou seja, pior a sua performance, ou maior a diferença entre as previsões e os valores reais.

Este algoritmo calcula (dentro de intervalos arbitrários definidos pelo pesquisador), utilizando validação cruzada, os melhores valores para o número de iterações, e também os melhores números: a) de camadas ocultas em uma NN; e b) de neurônios por camada, de forma a obter o melhor valor para a métrica selecionada para a avaliação da performance da ANN.

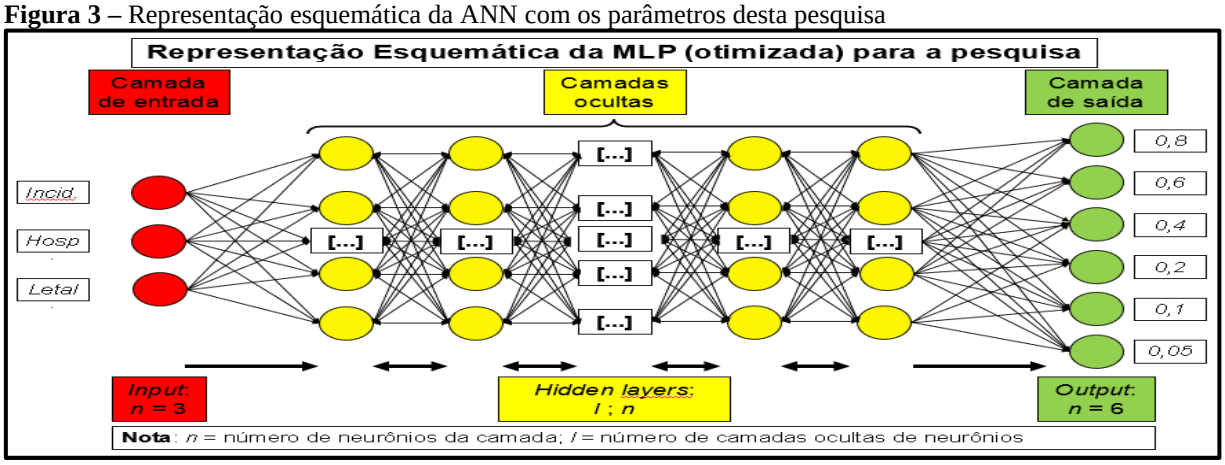
A Fig. 3 mostra esquematicamente como é o design da ANN projetada para esta pesquisa específica, e que usa as variáveis e parâmetros desta pesquisa. Este design inclui:

a) três neurônios na camada de entrada – pois foram selecionadas três variáveis preditoras: a incidência e as taxas de hospitalização e de letalidade da COVID-19;

b) um número (a ser definido durante a execução do algoritmo) de camadas ocultas de neurônios e de iterações, para a obtenção do menor MSE, visando otimizar as predições;

c) seis neurônios de saída – correspondendo às seis classes (ou intervalos de variação percentual nos fluxos (financeiros e de pacientes) dos consultórios), como mensuração dos efeitos econômicos da pandemia nestes estabelecimentos.

Figura 3 – Representação esquemática da ANN com os parâmetros desta pesquisa



3.8.9 Algoritmo de Machine Learning Automatizada (AutoML)

Conforme descrito nas seções anteriores, um processo de ML – para a análise de problemas que identifiquem associações menos evidentes – pode assumir configurações de complexidade mediana a bastante elevada. E, visando a simplificação das configurações dos workflows correspondentes a estes processos, foi desenvolvido pela equipe do Knime o *component* denominado Automated Machine Learning (ou *AutoML*). Nas *webpages* relativas a ele, é apresentada e esmiuçada a sua execução com a incorporação de 10 (ou outro número arbitrário de) diferentes algoritmos de ML supervisionados tanto para classificação binária quanto para a multiclasse. Neste *AutoML*, há pouca (se alguma) demanda de que o usuário faça manualmente as configurações para o seu funcionamento, uma vez que o próprio *component* já foi desenvolvido sob o objetivo de simplificar tanto quanto possível a rotina de programação

pelo “usuário comum” (não profissional de Tecnologia da Informação (TI)). Deste modo, o *Integrated Deployment* executa desde a preparação dos dados, passando pelo treinamento dos algoritmos, pela seleção e teste de parâmetros, pela apresentação dos dados de teste aos algoritmos treinados (incluindo a validação cruzada), pela avaliação de suas performances e até a seleção do algoritmo que apresentar os melhores desempenhos. Este *component* pode automatizar até mesmo todo o ciclo de ML, fazer uma parcela do pré-processamento dos dados, otimizar os parâmetros com validação cruzada, calcular os *scores*, avaliar o desempenho e selecionar os melhores parâmetros e resultados. E tudo com um mínimo de esforços para a configuração dos elementos deste workflow. O *component* também captura integralmente o processo, e oferece as saídas da implementação usando a *KNIME Integrated Deployment Extension* (ou Implementação Integrada do Knime). Este desenvolvimento foi feito juntamente com um tutorial, facilitando seu uso sem o treinamento formal em linguagens de programação, o que atende a um dos objetivos desta pesquisa.

A pesquisa aqui apresentada fez a comparação entre os algoritmos *k*-NN, SVM, Naïve Bayes, a rede neural MLP e o processo de ML automatizada (o AutoML).

3.8.10 Avaliação da performance dos algoritmos de classificação

A performance dos algoritmos de classificação selecionados durante a elaboração do modelo aqui apresentado foi avaliada comparativamente através dos parâmetros estatísticos para a obtenção da maior acurácia e da análise visual da Curva ROC (*Receiver Operating Characteristic*). A análise visual da curva ROC (KRZANOWSKI; HAND, 2009) foi originalmente desenvolvida para fins militares, durante a IIª Guerra Mundial, e menos de duas décadas depois passou a ser gradualmente incorporada às pesquisas em diferentes áreas da saúde. A abordagem básica para as curvas ROC envolve a comparação em classificações de objetos (ou pontos) em uma dentre duas possíveis classes. Ou mesmo simultaneamente (*i.e.*, no mesmo gráfico), entre diferentes pares de variáveis em que uma delas, a dependente, seja comum a todos os pares, sendo as demais determinantes dela. Assim, tem-se a comparação entre: a) as taxas de acerto na classificação de um dado ponto (mensurado pelos valores da variável em que ele é descrito) em uma de duas classes, *e.g.*, os que recebem a designação para pertencer a uma classe e efetivamente deveriam estar enquadrados nela (o que representa a taxa de Verdadeiros Positivos (TP)); e as taxas de erro neste enquadramento, em que um dado ponto foi indevidamente enquadrado nesta classe, porém deveria estar em outra (a taxa de Falsos

Positivos (FP)). Esta é a aplicação mais conhecida (e frequente) das Curvas ROC. Cabe recordar que, conforme pode ser visualizado no questionário desenvolvido para a pesquisa aqui apresentada, as respostas para as métricas dos consultórios dispõem de seis diferentes classes dentre as quais os participantes devem escolher uma que represente o quanto os seus fluxos foram afetados durante o período da pesquisa. Desta forma, para cada classe que recebe o enquadramento, há outras cinco classes que deixam de recebê-lo. Assim, a análise, por si só, já apresenta uma complexidade a mais para englobar todas as possibilidades referentes a esta escolha. No âmbito da DS, as ferramentas de análise que usam a Curva ROC tipicamente comparam as taxas de respostas de classificação que efetivamente correspondem a somente uma dentre somente duas opções (uma com valor "Positivo", e outra, "Negativo", sendo que ambas podem ser ou verdadeiras ou falsas). Para conduzir adequadamente esta maior complexidade, podem ser aplicadas diferentes alternativas. A adotada aqui foi a de selecionar iterativamente cada um dos possíveis valores da coluna "Classe" (*i.e.*, o valor, ou um rótulo) de cada uma das faixas de variação percentual nos fluxos), e calcular a Curva ROC para a comparação entre as respectivas taxas de TP e de FP. Assim, têm-se seis curvas para a série de seis valores desta variável categórica ordinal que representa cada uma das métricas. E então fez-se a comparação entre estas diferentes curvas, para extrair a informação desejada para quais das três taxas – incidência, de hospitalização ou de letalidade – representam as variáveis supostamente independentes (e das quais a dependente seja função), ou o quanto os fatores determinantes das reduções nos fluxos influenciam as variações nos fluxos mais do que o acaso. Em um gráfico de Curva ROC, as curvas que mais se aproximam do topo superior esquerdo são as que representam associações maiores do que ao acaso – uma vez que a reta diagonal que vai do canto inferior esquerdo até o topo superior direito representa a relação TP/FP com valores determinados pelo acaso, e que os pontos traçados abaixo desta reta representam pontos mal classificados, *i.e.*, posicionamentos para os quais a designação de que as variações observadas naquela faixa não seriam associados à variável tida como supostamente determinante (ou independente) – e assim a análise será feita para identificar se, para uma dada métrica e com um dado valor para a resposta observada, uma (ou mais) das três taxas podem, de fato, estar associadas à respectiva observação.

3.9 ELABORAÇÃO DO RELATÓRIO DE PESQUISA SEGUNDO OS STARE-HI

Esta pesquisa terá sua redação readequada, onde possível, ao Statement on reporting of evaluation studies in Health Informatics (STARE-HI). Estas diretrizes visam a padronização em publicações de estudos para a avaliação de aplicativos em Informática para a Saúde (BRENDER *et al*., 2013). Estes "princípios" serão aplicados na redação deste relatório. Dada a sua importância, é crítico avaliar tanto a robustez quanto a otimização dos sistemas informatizados aplicáveis a dada realidade prática na gestão e entrega de serviços de saúde, em cada conjunto peculiar de variáveis e valores.

Para as diferentes áreas e tipos de investigação científica na saúde, já foram propostos diversos sistemas de diretrizes (e outras orientações), *e.g.*, CONSORT, SPIRIT, CARE, STARD, AGREE, STROBE, PRISMA, CHEERS, etc., disponíveis no website da rede EQUATOR. O STARE-HI foi proposto para aperfeiçoar este conjunto de diretrizes, especificamente nos estudos de avaliação de estudos de sistemas de saúde. Devem ser usados como diretrizes principais ou, no caso de o modelo de estudo já ser contemplado por outro conjunto de diretrizes, servir como complementares a estas.

4 RESULTADOS

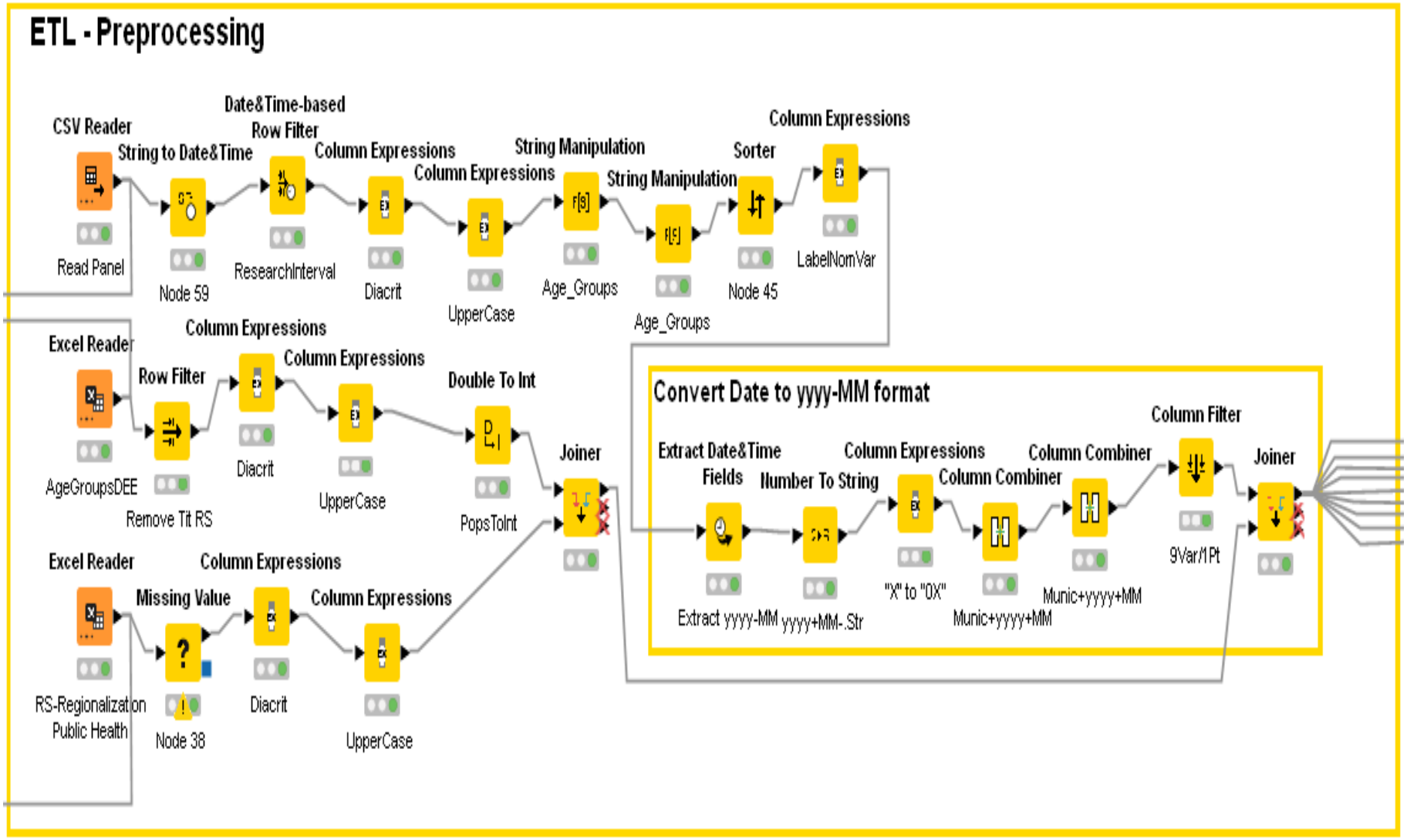
Este capítulo apresenta os resultados sistematizados da pesquisa para o desenvolvimento do modelo pretendido para a aplicação de ferramentas de DS, conforme a divisão resumida na Seção 3.2, começando pela descrição de um panorama geral da COVID-19 no RS e das relações identificadas entre as variáveis selecionadas. Subsequentemente, são representadas as associações entre as variáveis que descrevem o quadro geral COVID-19 e as métricas dos fluxos de pacientes (e financeiros) dos consultórios, conforme respostas coletadas na *survey*. Ao final desta seção, são apresentados os workflows e os resultados dos algoritmos usados para a construção de um modelo para a predição da flutuação nas métricas em função das variações nas taxas usadas para avaliar a disseminação e gravidade da COVID-19 no RS. Exceto quando explicitamente indicado de outro modo, todos os workflows e demais ilustrações (tabelas, gráficos e figuras) foram gerados no Knime e se referem aos casos de COVID-19 registrados no RS e/ou às respostas à *survey*, analisados durante a pesquisa feita nos primeiros 16 meses da pandemia (*i.e.*, março de 2020 a junho de 2021).

O primeiro dos resultados, e também a ferramenta para a obtenção de todos os demais, é o pré-processamento dos dados coletados, que foi feito seguindo uma sequência de etapas relativamente extensa e já mencionada no capítulo precedente, para viabilizar as análises subsequentes. Uma visão generalizada do workflow construído ao longo da pesquisa, em toda a sua complexidade é mostrada no Apêndice C. Em cada seção é apresentado e detalhado o trecho do workflow correspondente a elas e, onde conveniente, são usados *metanodes* ou *components*, para facilitar a visualização e compreensão das atividades de processamento, sem deixar a área sobrecarregada de informações pela multiplicidade de *nodes* envolvidos.

4.1 PRÉ-PROCESSAMENTO DOS DADOS DA COVID-19 (ETL)

A primeira etapa deste pré-processamento é designada de “Extração, Transformação e Carregamento dos dados” (ETL). O processo de KDD inicia com a ETL dos dados para a ferramenta em que serão analisados. Os dados foram extraídos e selecionados de diferentes fontes, foi feita uma uniformização de seus formatos, tamanho, cor, maiúsculas, remoção de diacríticos, filtragem por intervalo entre datas e unidades de valores, com a seleção dos dados conforme os interesses da pesquisa. O trecho do workflow para o ETL é visualizado na Fig. 4. A razão para a não inclusão da DB resultante neste trabalho é o seu tamanho absoluto, com dados de 1,325 milhão de casos, hospitalizações e óbitos devidos à doença.

Figura 4 – Workflow: ETL dos dados da COVID-19



Nos *nodes* iniciais, estão os WIMPs para a importação para o Knime dos arquivos (ou de outras fontes externas, painéis, sites ou outras fontes online, com acesso restrito ou aberto, conforme fosse o caso) usados no projeto. E os *nodes* subsequentes contêm as funcionalidades para cada uma das atividades sequenciais deste processo de ETL, de forma a uniformizar e pré-processar todos os dados para as etapas subsequentes de ML propriamente ditas.

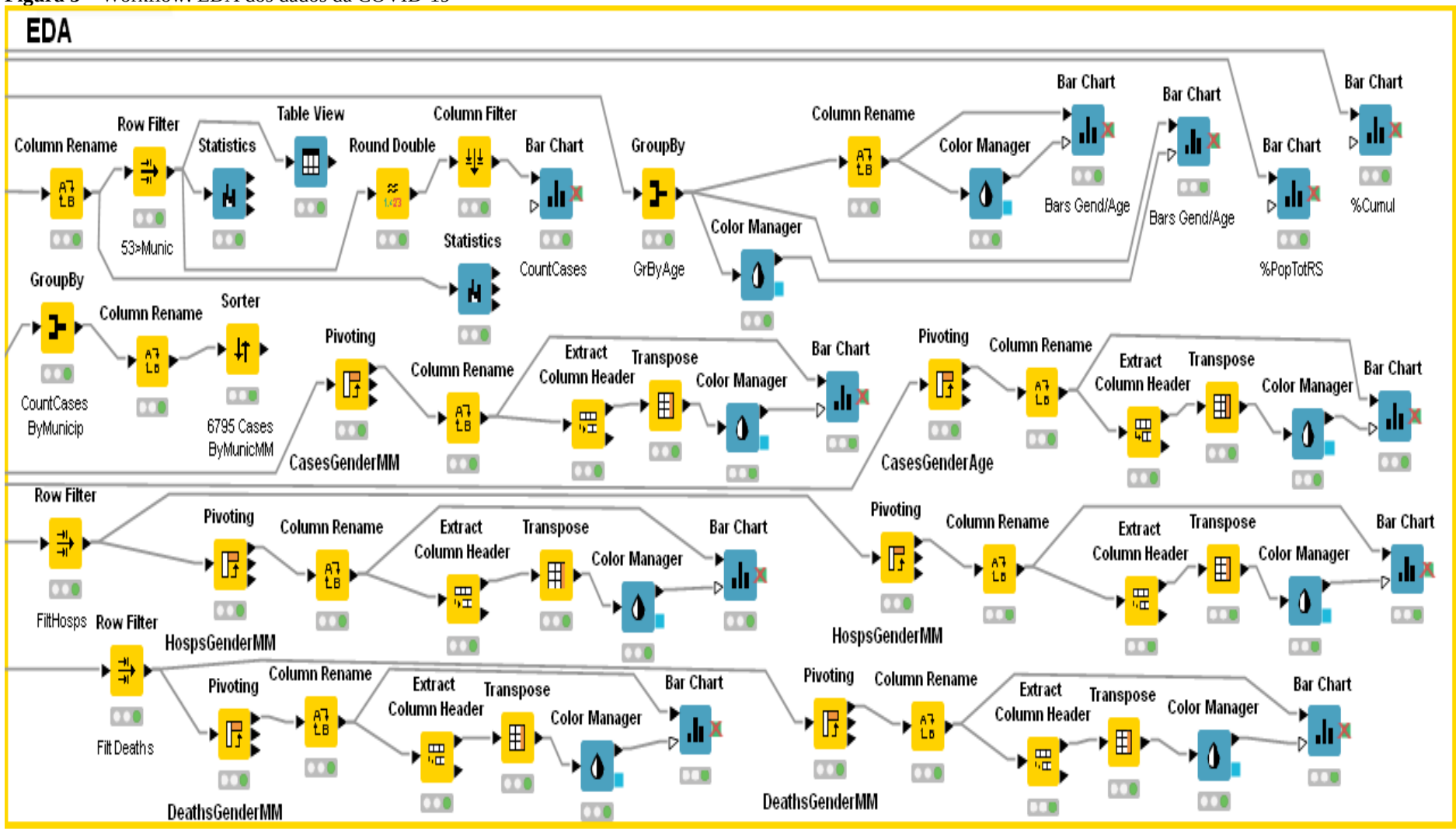
4.2 ANÁLISE EXPLORATÓRIA DOS DADOS DA COVID-19 (EDA)

Como já comentado anteriormente, esta seção (e suas subdivisões) mostram a Análise Exploratória dos Dados (EDA), que geralmente representam uma abordagem inicial feita para a maior familiarização do pesquisador com os dados e com a realidade que representam. As etapas e procedimentos de análise envolvidos na EDA estão mostrados graficamente na Fig. 3 e são descritos em maior detalhe nas seções subsequentes.

Esta etapa prática dos trabalhos permitiu o início de todas as análises posteriores, e serviu também como ferramenta para as principais análises dos perfis epidemiológicos da COVID-19 no RS. Este aprofundamento permitiu que fosse compreendido não somente a aplicação de um espaço *n*-dimensional a um problema concreto de pesquisa. No caso presente, foram usadas três dimensões para a composição matemática de uma visão de um espaço em que três diferentes variáveis (as supostas “variáveis independentes”) descrevessem diferentes aspectos (ou “dimensões”) deste problema de pesquisa específico. Representações análogas poderiam ser facilmente desenvolvidas para quaisquer outros números de variáveis (ou dimensões) selecionadas para analisar outras situações, e mesmo com outros conjuntos de dados (ou valores de tais variáveis), e com a simplicidade conceitual desejada para cada situação.

A construção do trecho correspondente do workflow ilustra como pode ser feito o encadeamento das diferentes operações sequenciais selecionadas para a análise de um conjunto de dados, visando que possa ser atingido um dado objetivo de pesquisa através de um raciocínio analítico desenvolvido para cada análise específica.

Figura 5 – Workflow: EDA dos dados da COVID-19



4.2.1 Distribuição da população nos diferentes municípios do RS

A presente seção mostra os resultados das atividades feitas para analisar o mapeamento de aspectos sociopolíticos e epidemiológicos da população do RS.

A população do RS, que é de quase 11,5 milhões (11.422.973) de habitantes nos diferentes municípios gaúchos, e tem uma distribuição marcadamente desigual entre os 497 diferentes municípios do RS. Os Gráfs. 1(a) e (b) mostram esta disparidade na distribuição, respectivamente para os percentuais municipais da população do RS e para a soma cumulativa destes percentuais para os 53 municípios mais populosos. A atual distribuição mostra que:

a) praticamente 1/8 (ou 12,72%) de toda a população reside na capital;
b) o segundo município mais populoso (Caxias do Sul) tem cerca de 1/3 do número acima, ou seja, 4,2% da população total do RS;
c) apenas dois outros municípios (Canoas e Pelotas) têm mais do que 300.000 hab.;
d) 22 municípios (ou menos de 5% dentre o número total dos 497 municípios gaúchos), concentram mais da metade de toda a população do Estado;
e) se forem computados os 53 municípios gaúchos mais populosos (ou mais do que o dobro do subtotal anterior, porém cerca de apenas 10% do número total de municípios) vê-se neles concentrados dois terços (66,7%) do total dos habitantes do RS, e também mais do que dois terços (68,3%) do total de casos registrados;
f) apenas 169 municípios (mais do que o triplo do subtotal anterior, porém apenas 1/3 do número total de municípios gaúchos) têm mais do que 10.000 hab.;
g) no extremo oposto da pirâmide da distribuição populacional, mais da metade (252, ou 50,07%) dos municípios têm menos do que 6.000 hab.

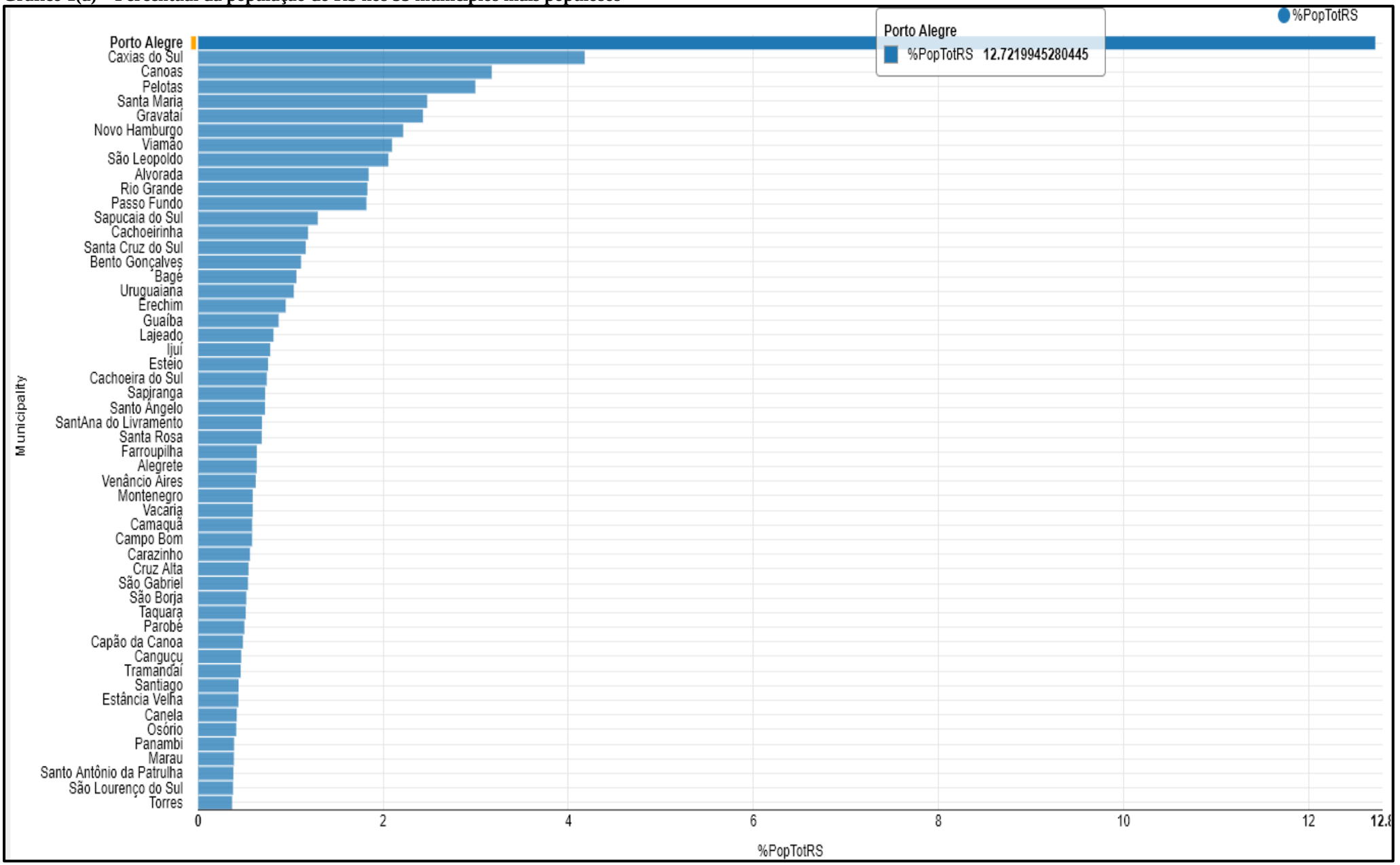
A pesquisa também investigou associações entre a densidade populacional e as métricas adotadas para avaliar a disseminação e gravidade da pandemia no RS. Conforme será revisto nas seções que apresentarão resultados e discutirão as possíveis associações entre números de habitantes e a disseminação da doença, esta desigualdade na distribuição das concentrações populacionais – ou antes, a própria concentração e circulação de habitantes em maior proximidade – tem um papel marcado na facilidade de contágio pelo SARS-CoV-2.

A análise dos perfis populacionais servirá de base para a investigação subsequente de eventuais associações entre totais (ou subgrupos) populacionais e respectivas incidências e taxas de hospitalização e de letalidade da COVID-19 no RS.

O significado e implicações das análises feitas para a distribuição da COVID-19 podem ser mais facilmente compreendidos através da visualização do percentual acumulado da população por município, o que reforça a grande desigualdade nesta distribuição, e sua participação ponderal nestes subtotais parciais. Os Gráfs. 1(a) e (b) ilustram a desigualdade de concentrações populacionais calculadas entre os 53 municípios mais populosos do RS. O Gráf. 1(a) mostra o percentual da população do RS que os diferentes municípios mais populosos representam. E o Gráf. 1(b) mostra que, à medida que se prossegue na lista decrescente das

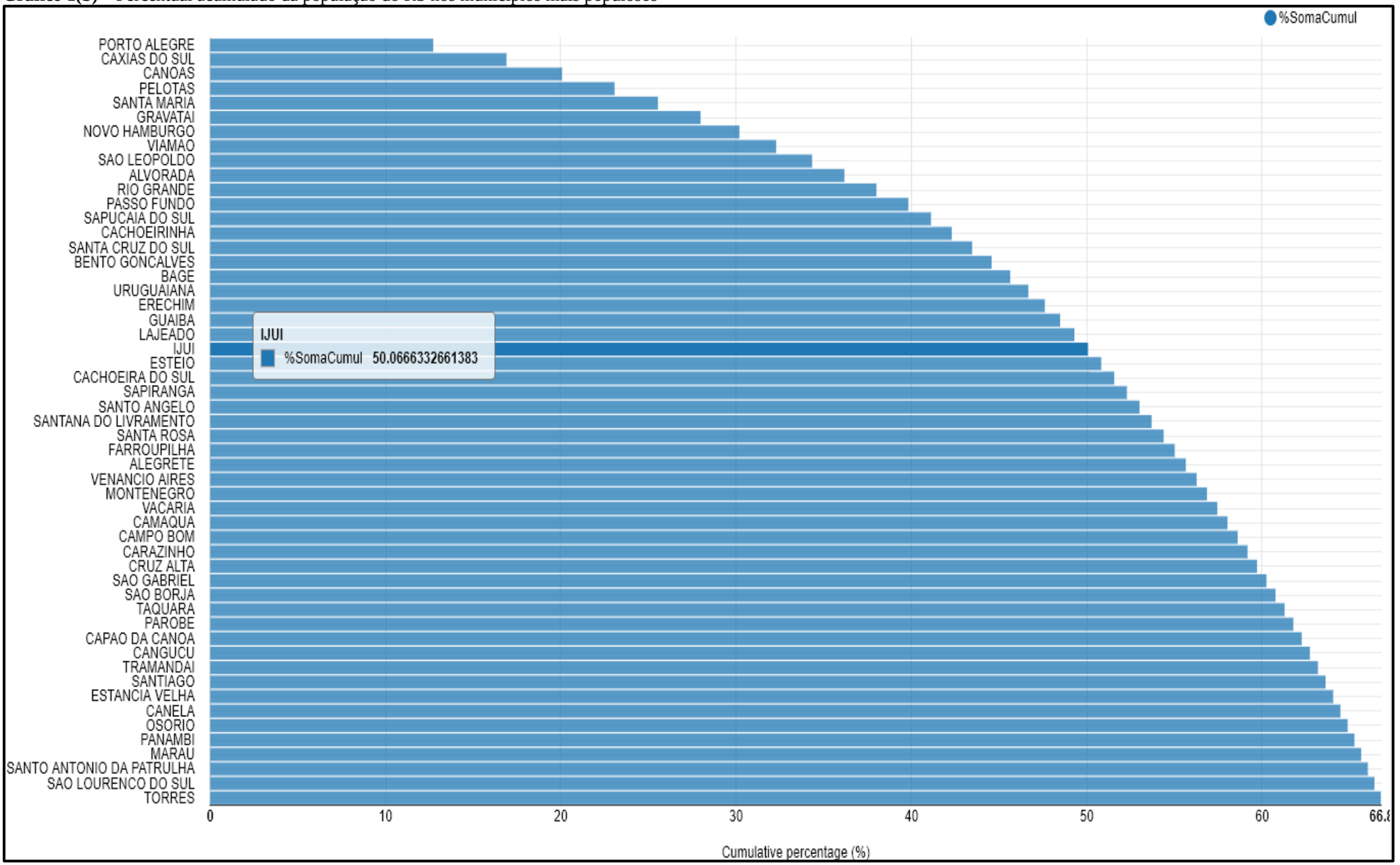
populações municipais, menor será o acréscimo percentual de cada população local. Neste gráfico, são destacados os 22 municípios que reúnem metade da população do Estado, e os 53 municípios com dois terços da população total do RS. Esta análise, parece sugerir a validade da investigação de uma (eventual) associação entre números populacionais locais e fatores de contágio pela doença, por transmissão sustentada (também dita “transmissão comunitária”) inter-humanos.

Gráfico 1(a) – Percentual da população do RS nos 53 municípios mais populosos



Nota: O município de Porto Alegre está destacado (em cor e por uma *flag*) apenas para ilustrar o recurso do *pointer*

Gráfico 1(b) – Percentual acumulado da população do RS nos municípios mais populosos



Nota: O destaque de Ijuí separa 22 municípios que somam 50% da população do Estado, dentre os 53 municípios com 2/3 da população total do RS.

4.2.2 Distribuição dos casos de COVID-19 no RS

Foi feito o pré-processamento da totalidade dos casos registrados no RS ao longo do período dos 16 meses iniciais da pandemia cobertos pela pesquisa, a contar de março de 2020, o que resultou em aproximadamente 1.325.000 (1.324.589) casos, 101.194 hospitalizações e 33.069 óbitos[7]. Nestes registros, foi aplicado o *node Statistics*, e dele foram extraídos os dados descritivos da população e casos, apresentados nas Tabelas 1(a)-(c):

a) gênero dos pacientes

Tabela 1(a) – Distribuição dos casos, conforme gênero (declarado) dos pacientes

Gênero dos pacientes	N° total de casos	Percentual do total (%)
Feminino	701.509	52,96
Masculino	623.080	47,04
Não Binário	0	0,00
Prefere não informar	0	0,00
Totais	1.324.589	100,00

b) distribuição dos casos conforme método de diagnóstico

Tabela 1(b) – Distribuição dos casos selecionados, conforme o método de diagnóstico

Método de diagnóstico	N° total de casos	Percentual do total (%)
Teste Rápido	622.482	46,99
RT-PCR	662.861	50,04
Subtotais Teste Rápido + RT-PCR	1.285.343	97,03
Ignorado	3	0,00
Outros Testes	23.220	1,75
Clínico	7.215	0,54
Clínico-Epidemiológico	3.738	0,28
Clínico-Imagem	5.070	0,38
Totais	1.324.589	100,00

Quanto ao método de confirmação dos diagnósticos, conforme a Tabela 1(b), quase a totalidade (97%) foi feita por RT-PCR (*Reverse Transcriptase – Polymerase Chain Reaction*) (50,04%) ou Teste Rápido (46,99%), ou seja, em proporções muito próximas (1,00 : 0,94).

[7] Até a data da elaboração deste relatório, o total investigado de casos (durante os 16 meses da pesquisa) corresponde a quase a metade (45,59%) do total de casos registrados desde o início da pandemia até o presente; enquanto as hospitalizações corresponderam a 77,4% (mais de três quartos); e os óbitos responderam por 79,6% (quase quatro quintos) do total registrado.

c) Distribuição mensal e por gênero dos casos de COVID-19

Tabela 1(c) – Distribuição por gênero dos casos mensais de COVID-19

Data de confirmação (yyyy-mm)	Casos/mês (F)	Casos/mês (M)	Casos/mês (Totais)
2020-03	265	254	519
2020-04	1.119	903	2.022
2020-05	4.778	4.281	9.059
2020-06	11.167	10.268	21.435
2020-07	27.714	25.281	52.995
2020-08	33.246	30.431	63.677
2020-09	27.034	24.320	51.354
2020-10	28.861	26.439	55.300
2020-11	57.503	51.096	108.599
2020-12	73.391	60.590	133.981
2021-01	51.481	41.770	93.251
2021-02	84.409	72.651	157.060
2021-03	121.862	110.677	232.539
2021-04	52.001	46.802	98.803
2021-05	67.123	62.190	129.313
2021-06	59.555	55.127	114.682
Totais (2020-03 a 2021-06)	701.509	623.080	1.324.589

4.2.3 Associação entre a população dos municípios e número de casos (NCases)

Visando dar suporte para avaliações sobre a transmissibilidade da doença, foi analisada a distribuição dos casos em função da população dos municípios em que estes casos ocorreram. Buscou-se investigar esta associação para esclarecer algumas das variações nos valores (brutos) assumidos pelas variáveis de trabalho, e para a posterior ponderação destes números através das taxas adotadas, conforme mencionado no Cap. 3. Foi feita a aplicação do *node* *Linear Correlation*, que permite o cálculo do coeficiente de correlação *r* de Pearson, e traz os resultados apresentados na Tab. 2:

Tabela 2 – Correlação linear entre a contagem de casos e a população de cada município

Row ID	S First colum...	S Second...	D Correlation value	D p value
Row0	CountCasesMunic	PopTotMunic	0.9896900543465...	0.0

Através da aplicação do *node* *Linear Regression Learner* – que já pode ser adotado mesmo na fase da EDA, sem ter-se que aguardar até as fases de ML do projeto para utilizá-lo – obtém-se a matriz desta regressão linear e o *scatter plot* dos dados relacionando ambas as variáveis, *i.e.*, a contagem de casos em função da população de cada municipalidade, conforme mostrado na Tab. 3, em que o R^2 (o quadrado do coeficiente da correlação) é 0,9795, e no Gráf. 2:, que mostra uma reta passando muito próximo da grande maioria dos pontos.

Tabela 3 – Regressão linear: população municipal *vs* casos de COVID-19

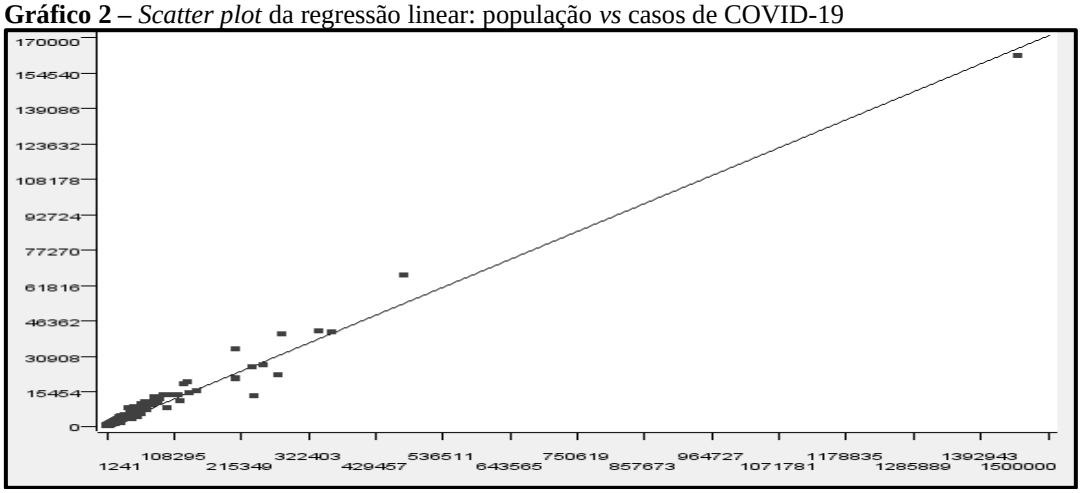
File

Statistics on Linear Regression

Variable	Coeff.	Std. Err.	t-value	P>|t|
PopTotMunic	0.1141	0.0007	153.7379	0.0
Intercept	52.3034	61.2159	0.8544	0.3933

R-Squared: 0.9795
Adjusted R-Squared: 0.9794

Gráfico 2 – *Scatter plot* da regressão linear: população *vs* casos de COVID-19



Legenda: População municipal (eixo *x*) em habitantes e contagem de casos (eixo *y*).

O resultado desta análise tem o valor do coeficiente de correlação (o que é a raiz quadrada do parâmetro R^2) de aproximadamente 0,99, um erro-padrão muito baixo (= 0,0007) e um *p*-valor calculado como $p = 0,0$. E isto significa 98,97% de associação, que é uma associação altamente significativa (e quase perfeita) entre o contingente populacional de cada município e o respectivo número de casos de COVID-19.

4.2.4 Comparação entre as variáveis SRAG e Hospitalização

Os casos sintomáticos graves foram registrados nos diferentes locais de atendimento, usando diferentes sinais e sintomas (em sua maioria através de variáveis binárias apresentadas na Seção sobre as DBs empregadas neste estudo), que foram abastecidas da DB do DataSUS, a qual foi usada para alimentar o painel da SES-RS para o enfrentamento da COVID-19. Dentre estas variáveis, duas em particular foram consideradas nesta seção: a) a dos registros como "SRAG" (ou com sintomas graves); e b) e a dos que foram hospitalizados. Aparentemente não houve, no momento de cada registro, uma distinção significativa entre ambas, *i.e.*, entre o enquadramento dos casos portadores da síndrome sintomática da COVID-19 (a SRAG) e os casos que levaram os pacientes a internações hospitalares. E, para analisar se esta distinção

poderia ter maiores efeitos sobre o objeto deste estudo, investigou-se esta possível associação usando novamente o *node* *Linear Correlation* do Knime (Gráfs. 3 e 4).

Gráfico 3 – *Scatter plot* de registros SRAG *vs* Hospitalizações por COVID-19

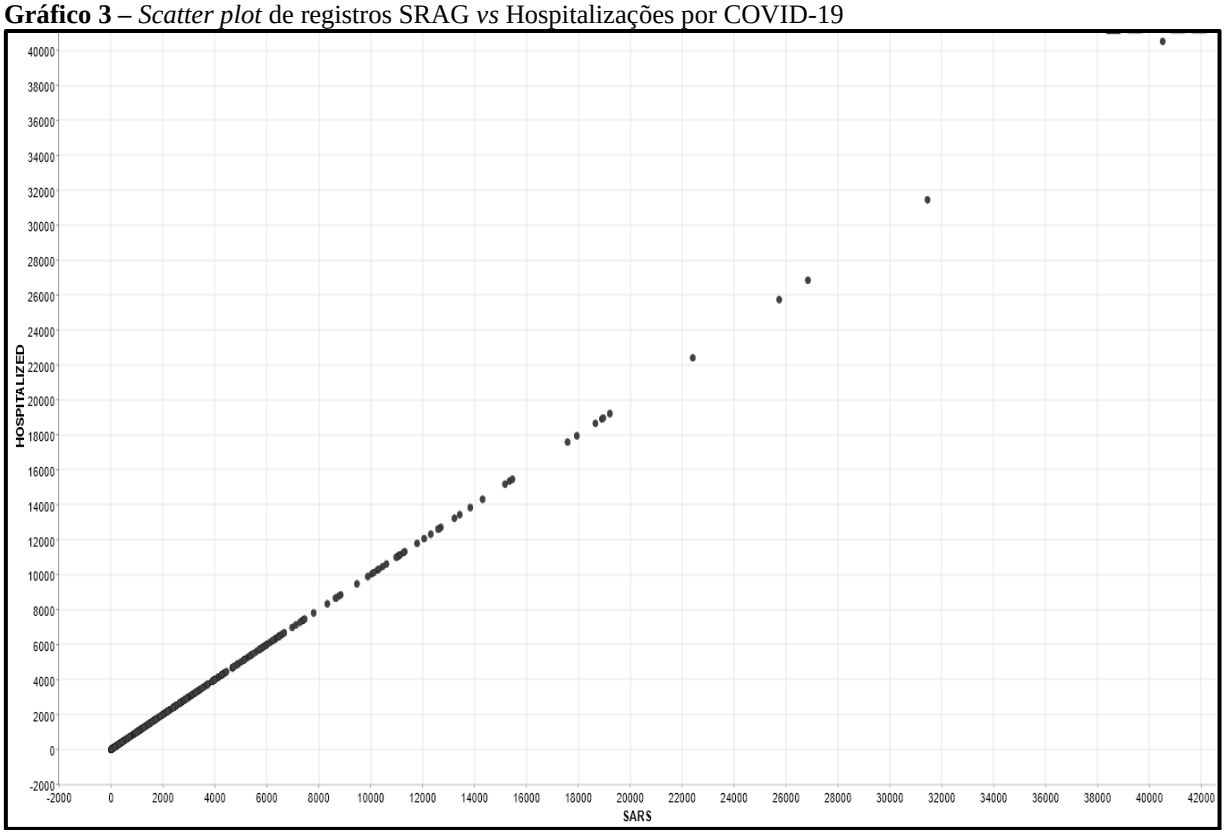
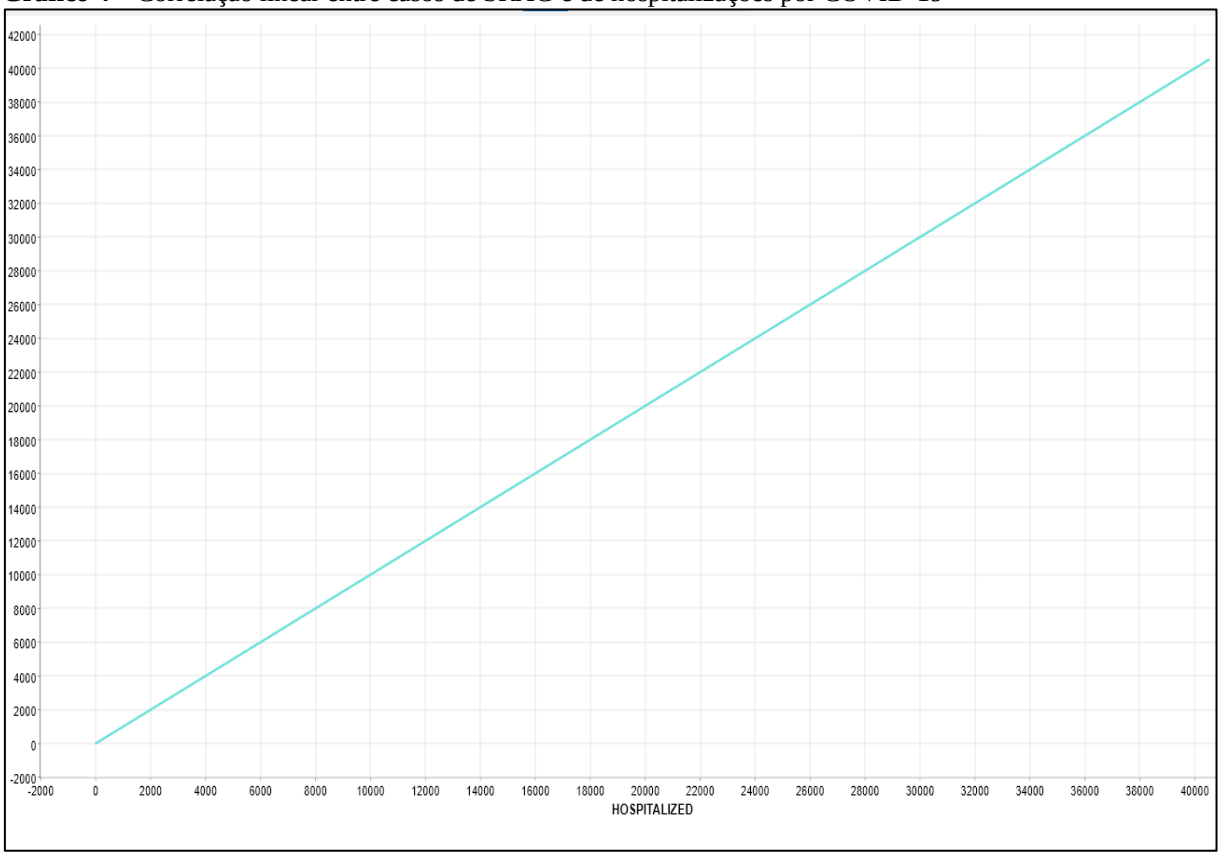


Gráfico 4 – Correlação linear entre casos de SRAG e de hospitalizações por COVID-19



Procedendo-se à análise da correlação linear, chega-se à matriz de correlação linear mostrada na Tab. 4:

Tabela 4 – Correlação linear entre as variáveis SRAG e Hospitalização

Row ID	S First col...	S Second...	D Correlation value	D p value
Row0	HOSPITALIZED	SARS	0.9939767923511...	0.0

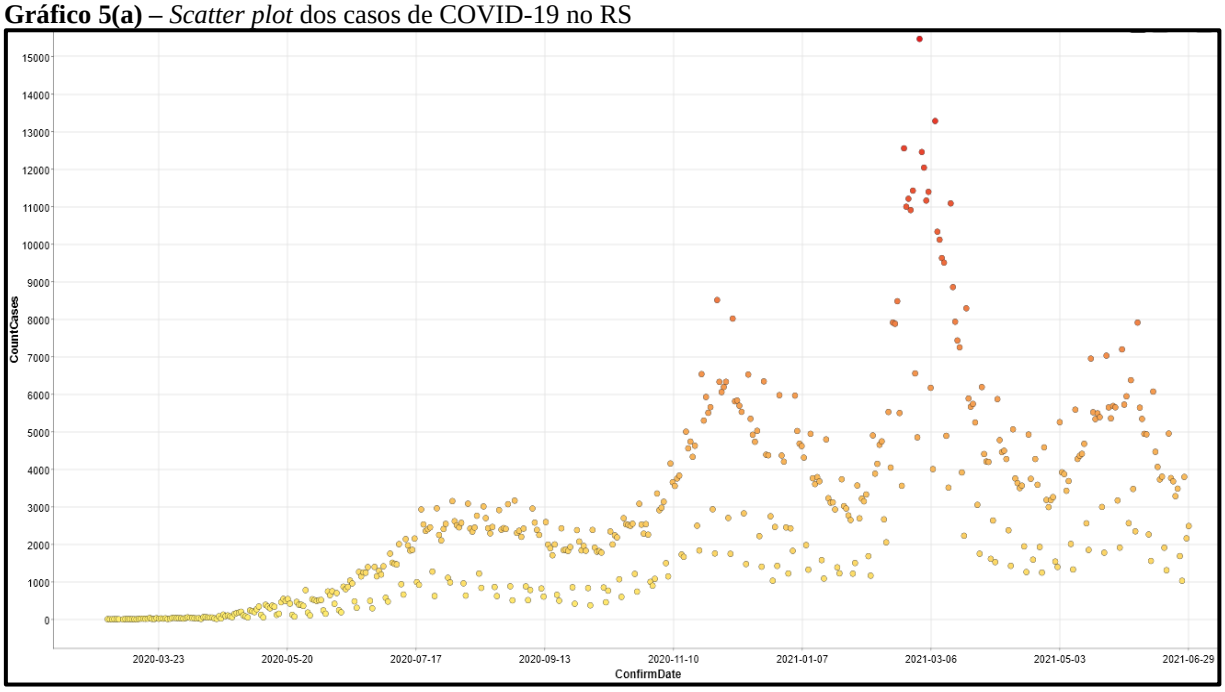
Esta análise fornece o valor da correlação de aproximadamente 0,994, e tem um *p*-valor calculado como $p = 0{,}0$ (para a precisão decimal usada no modelo). E isto significa uma associação altamente significativa (e praticamente perfeita) entre os pacientes que foram hospitalizados e os casos de registros com SRAG.

A partir das representações e cálculos acima, ao longo do presente trabalho, as variáveis categóricas binárias “SRAG” e “Hospitalizações” foram tomadas como equivalentes, para todas as finalidades desta pesquisa, sendo ambas usadas indistintamente para avaliar a severidade (não extrema) dos casos.

4.2.5 Distribuição temporal dos casos, hospitalizações e óbitos

Os casos de COVID-19 no RS durante o período da pesquisa, em um total de 1.324.589 casos registrados, estão distribuídos ao longo do intervalo da pesquisa e, a seguir, por faixa etária e gênero, conforme Gráfs. 5(a)-(c):

Gráfico 5(a) – *Scatter plot* dos casos de COVID-19 no RS



O Gráf. 5(a) mostra os registros individualizados dos casos. O Gráf. 5(b) mostra a distribuição com o conhecido “padrão em ondas” das flutuações nos números de novos casos diários ao longo do tempo (ou, mais especificamente, ao longo do intervalo coberto pela pesquisa), bem como as diferentes magnitudes de cada “onda”. Para facilitar a identificação deste padrão, foi aplicado o recurso da suavização (*smoothing*) desta curva, com o uso de médias móveis (MA), seguindo-se aqui o mesmo padrão de intervalo usado no DataSUS e no painel Covid da SES-RS, de 14 dias, conforme o Gráf. 5(c):

Gráfico 5(b) – Gráfico temporal dos casos de COVID-19 no RS

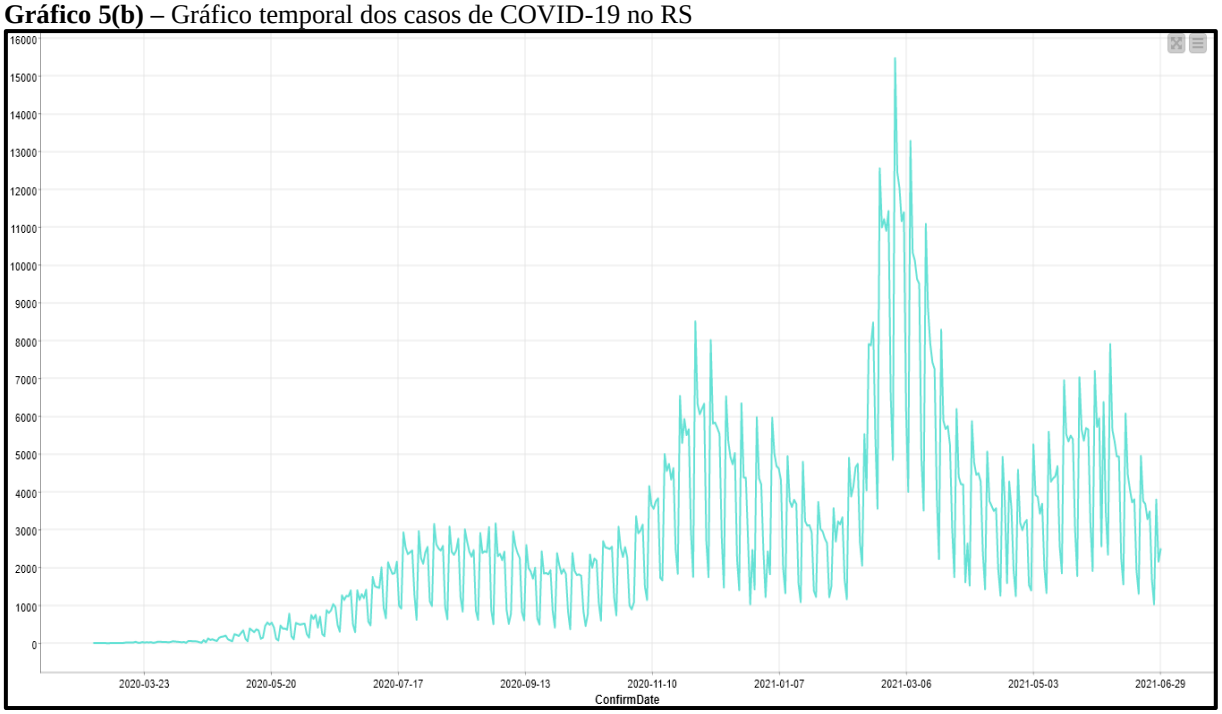
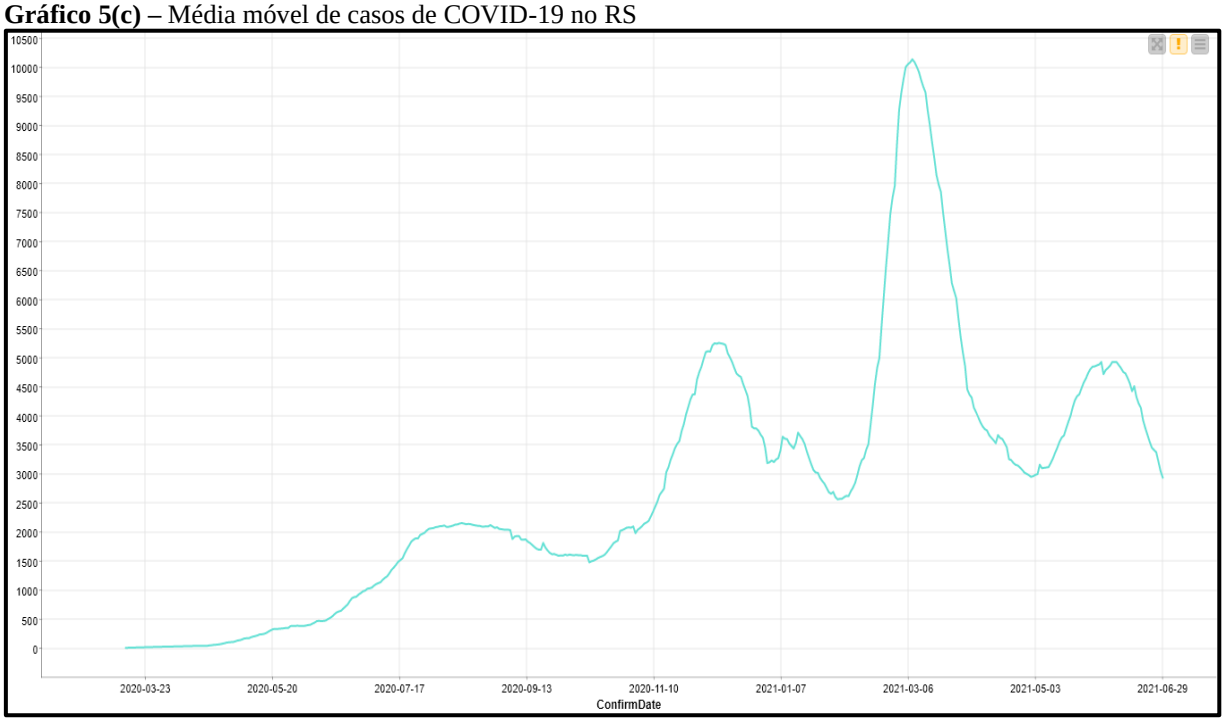
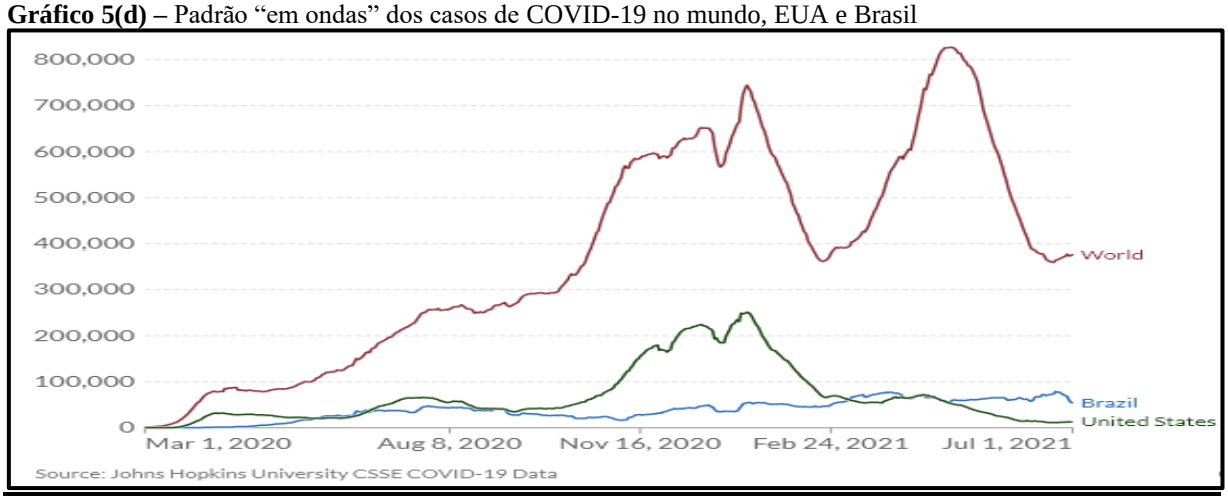


Gráfico 5(c) – Média móvel de casos de COVID-19 no RS



Para finalidades de comparação entre as “ondas” (de disseminação de novos casos de COVID-19) observadas no RS e as flutuações no Brasil, nos EUA e no mundo, durante o mesmo período da pesquisa descrita neste relatório, resgata-se aqui uma curva similar, gerada no painel da Johns Hopkins University (Gráf. 5(d)).

Gráfico 5(d) – Padrão “em ondas” dos casos de COVID-19 no mundo, EUA e Brasil



Fonte: COVID-19 Dashboard by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (JHU)

De modo similar ao uso dos totais e MA de casos, para avaliar a disseminação, a severidade (*i.e.*, registros de casos com hospitalização) e a severidade extrema (*i.e.*, registros de óbitos) da doença no período analisado, foi tomada pela equipe de pesquisa a decisão de parear a granularidade das dimensões de trabalho entre o panorama geral da COVID-19 e as respostas à *survey* sobre o mercado-alvo da pesquisa, mesmo sob a consciência de alguma perda de informações pelo nível de detalhe (LOD), devido a esta mudança na granularidade. Portanto, a partir desta seção, as ilustrações, cálculos, e distribuições correspondentes são agrupados por “mês-ano” como padrão de trabalho com os dados para as métricas, conforme Gráfs. 6(a)-(f).

As visualizações da distribuição, severidade e severidade extrema exploram a conjuntura ou panorama destes casos, para melhor embasar as análises quanto à suposição de que a pandemia possa ter desencadeado algum temor mais generalizado nas populações locais, quanto a possíveis riscos de frequentar consultórios odontológicos privados, ao menos em relação a procedimentos não urgentes, *i.e.*, os que poderiam ser postergados para após o período da pandemia, pois esta sempre foi considerada como temporária (e não permanente), e que, após sua vigência, a sociedade (e os serviços nela prestados) retomariam seus ritmos como anteriormente, supostamente em novo nível de equilíbrio dinâmico, tal como sucedeu em outras situações passadas, que desencadearam retrações temporárias no mercado.

Gráfico 6(a) – Disseminação da COVID-19 no RS, n° de casos por gênero e mês

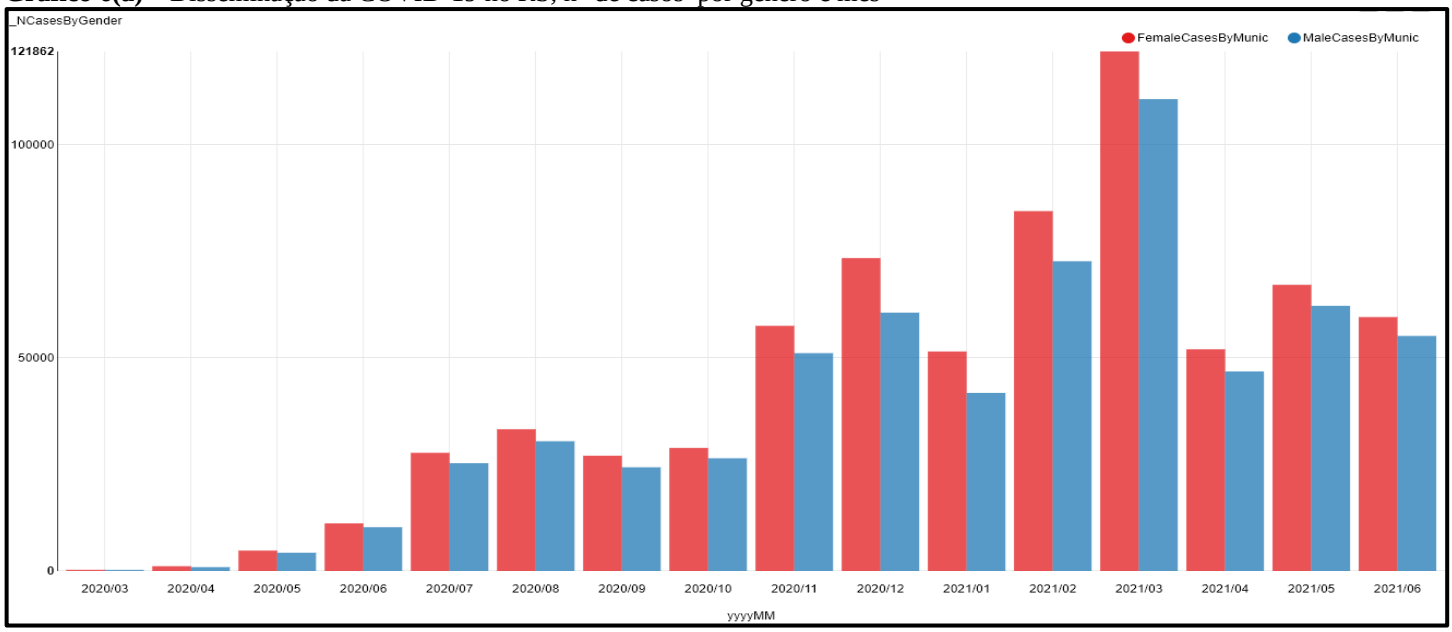


Gráfico 6(b) – Severidade da COVID-19 no RS, n° de casos por gênero e mês

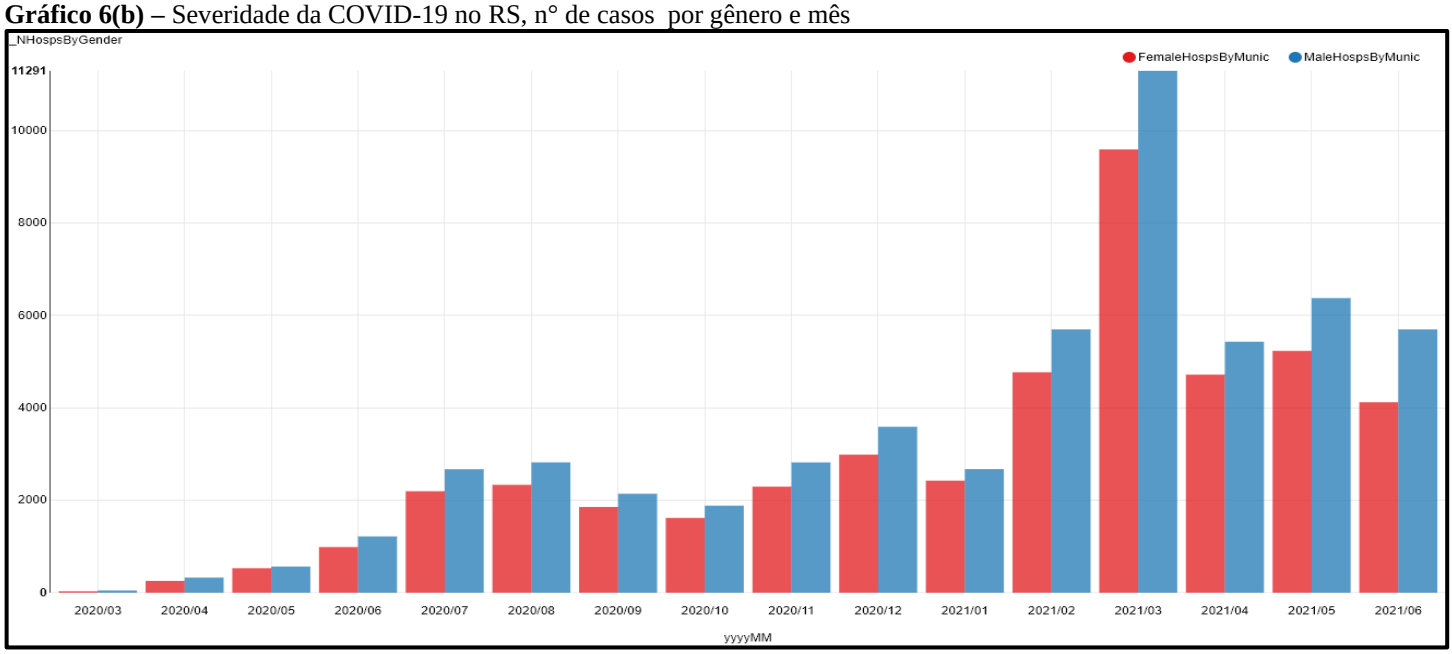


Gráfico 6(c) – Severidade extrema da COVID-19 no RS, n° de casos por gênero e mês

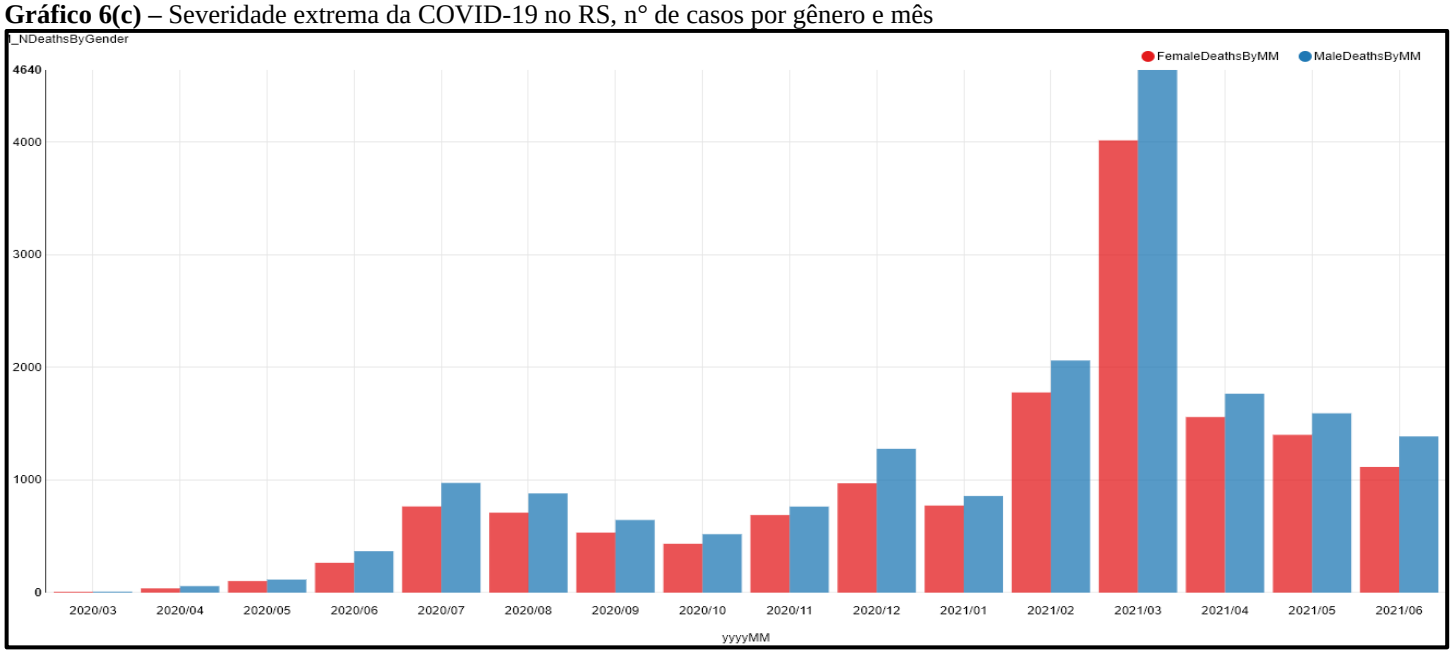


Gráfico 6(d) – Disseminação de COVID-19 no RS, n° de casos por gênero e faixa etária

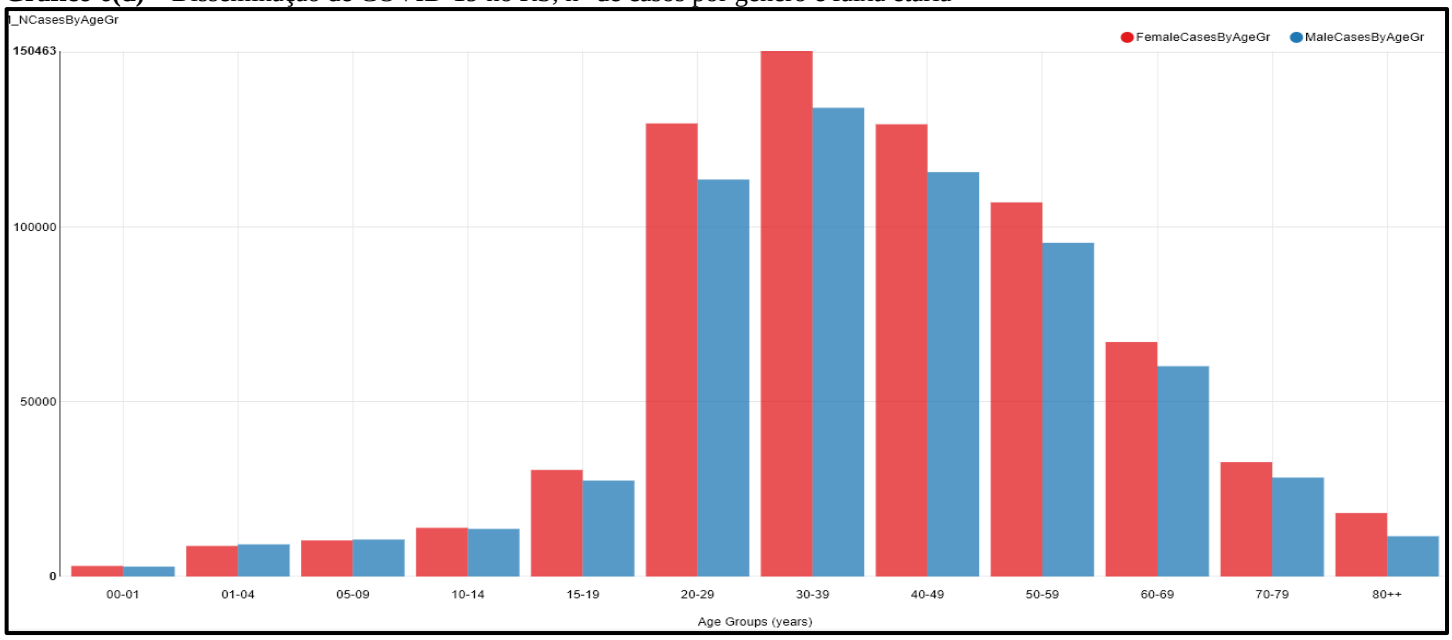


Gráfico 6(e) – Severidade da COVID-19 no RS, n° de casos por gênero e faixa etária

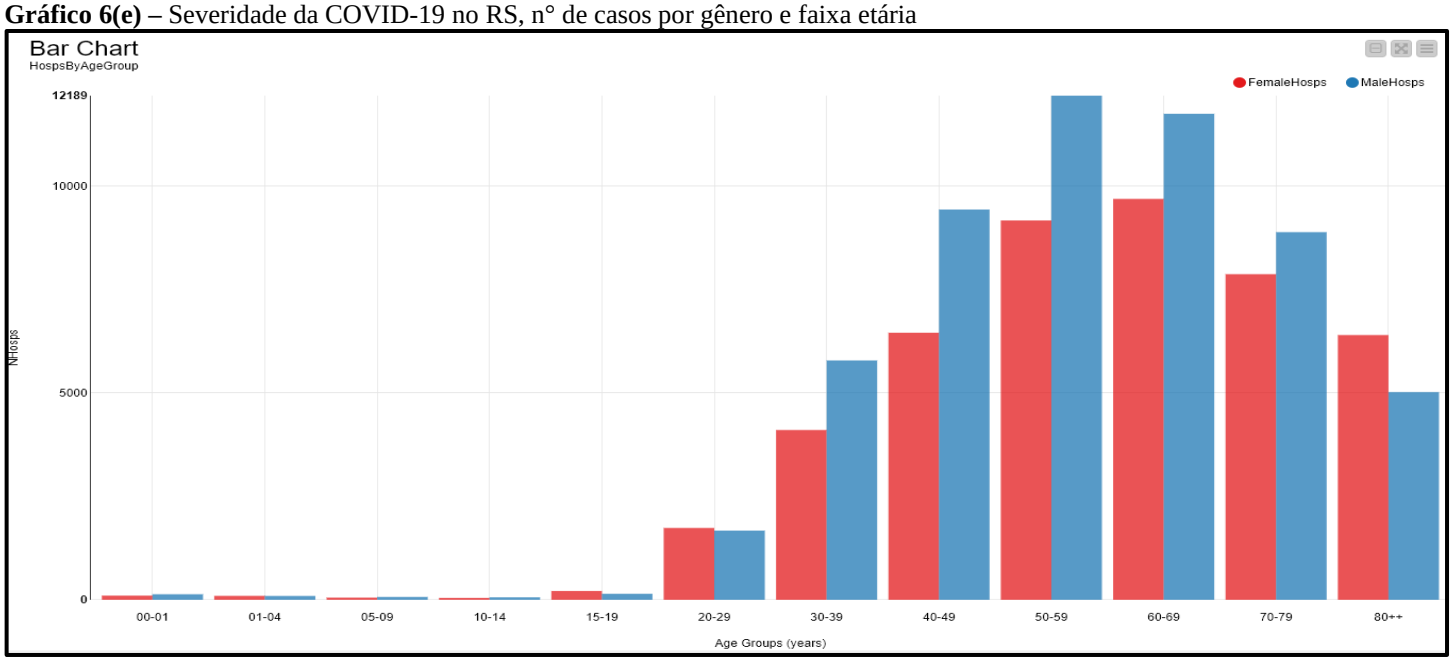
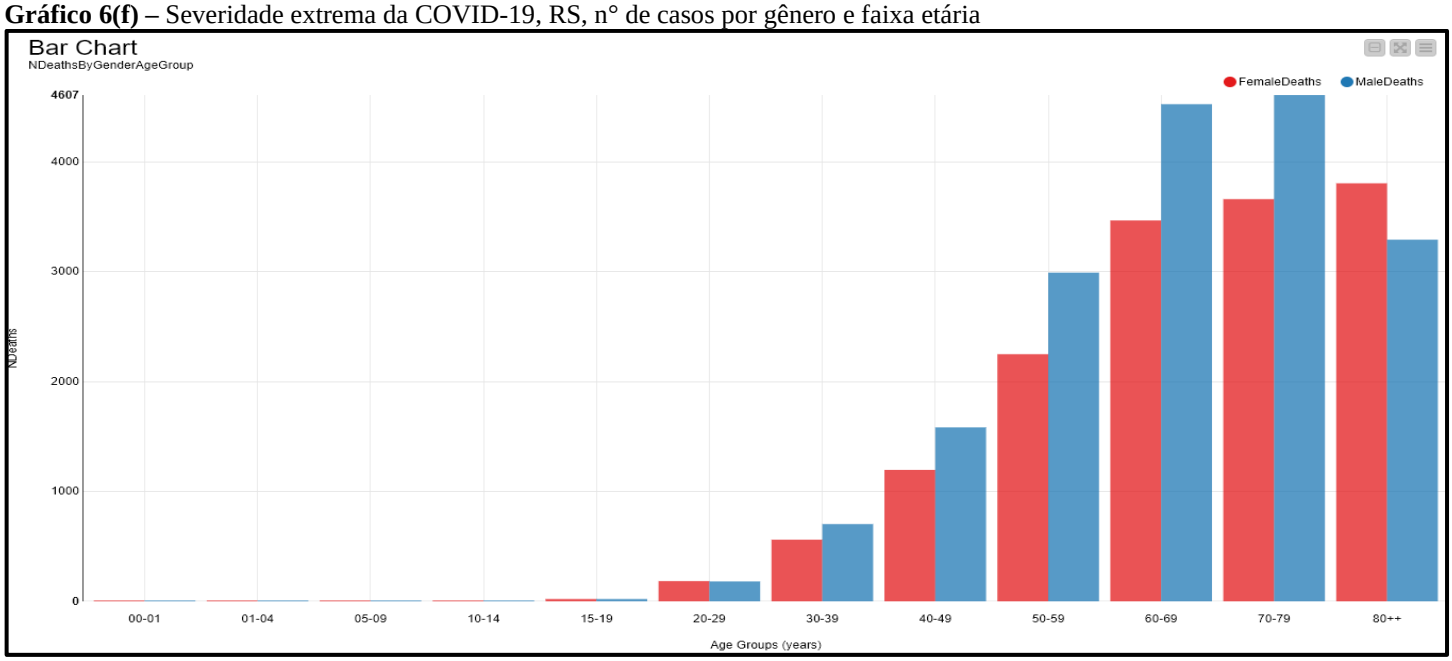


Gráfico 6(f) – Severidade extrema da COVID-19, RS, n° de casos por gênero e faixa etária



As representações acima destacam uma relativa inversão na distribuição anterior, com um aumento pequeno, porém consistente, das hospitalizações no gênero masculino, em relação ao feminino, em todo o tempo e (praticamente) em todas as faixas etárias. Esta observação permite constatar outra característica da doença, atingindo com maior gravidade de sintomas o gênero masculino.

As distribuições temporais e por faixa etária acima mostram uma diferença pequena, porém regular consistentemente maior de casos no gênero feminino e de hospitalizações e óbitos no gênero masculino, ao longo de todos os meses do intervalo da pesquisa e em todas as faixas etárias, repetindo regularmente o padrão "em ondas" característico da doença.

4.3 COMPARAÇÃO ENTRE DISTRIBUIÇÃO E SEVERIDADE

A partir das desigualdades mostradas nas visualizações anteriores – ao longo do tempo e por gênero – parece ser interessante visualizar os dados de outra maneira, de forma a tornar diretamente comparáveis os municípios com diferentes portes populacionais. Para tornar comparáveis entre si os dados da disseminação e gravidade da doença entre municípios de diferentes portes, a partir desta seção, o trabalho aqui apresentado adota o uso das três taxas.

Foram aplicadas estas taxas (OPAS, 2010), em lugar dos dados absolutos dos elementos a serem agrupados (ou clusterizados), aplicando transformações estatísticas aos dados para obter parâmetros pelos quais municípios de diferentes portes, agora diretamente comparáveis entre si, permitirão que sejam minimizadas eventuais distorções sobre o resultado da aplicação dos algoritmos. Com este objetivo, os dados foram transformados por meio da Normalização, reduzindo os diferentes "pesos" que maiores valores absolutos poderiam ter durante a aplicação dos algoritmos. Após a aplicação de cada algoritmo, os valores são "desnormalizados" (*i.e.*, devolvidos à sua escala original), para prosseguir a análise dos padrões identificados. O Knime tem *nodes* específicos para estas duas tarefas.

Ao longo de todo o presente trabalho, exceto onde indicado de outro modo, os dados são agrupados mensalmente, conforme uma (dentre duas) opções:

a) a divisão político-administrativa da Gestão em Saúde da SES-RS (Anexo A). O reagrupamento dos municípios (e Regiões de saúde) nas 21 "Regiões Covid" arbitradas pela SES-RS. Este agrupamento é baseado em critérios de proximidade geográfica, em um padrão que já é oficialmente adotado para os casos de COVID-19 e os diferentes municípios em que estes ocorrem, com objetivos de um melhor mapeamento e definições de políticas de enfrentamento à crise de saúde gerada pela pandemia, na expectativa de que levem à obtenção de melhores resultados de gestão, mais adequados às realidades destes locais, admitindo-se maiores semelhanças devidas à maior proximidade física (para estas, se aplica a anteriormente citada "Primeira Lei da Geografia"); ou
b) conjuntos dos municípios com taxas (mensais) semelhantes. Observou-se que, ao longo dos meses, as taxas variaram de maneira assíncrona entre os diferentes municípios. Este agrupamento foi feito em função dos objetivos de desenvolvimento de o modelo aqui proposto, para que este seja flexível e versátil, *i.e.*, que

possa ser mais facilmente adaptável a outras questões de pesquisa e a outras massas de dados a serem trabalhados.

O agrupamento em (a), *i.e.*, por proximidade geográfica, foi adotado como um *Gold Standard*, o padrão oficial já aplicado e reconhecido. E serviu para a comparação entre a qualidade dos resultados obtidos nesta pesquisa, através deste e de outros agrupamentos, feitos com base em (b), *i.e.*, de acordo com as taxas municipais mensais da COVID-19, através da aplicação dos algoritmos selecionados para o reagrupamento e extração de informações de uma massa de dados. Este reagrupamento (b) – feito em função de dados locais e com os recursos de classificação e de predição possibilitados por estratégias de ML – foi elaborado nesta pesquisa como possível estratégia alternativa à do agrupamento em (a).

4.3.1 Correlação entre Incidência e Taxas de Hospitalização e Letalidade

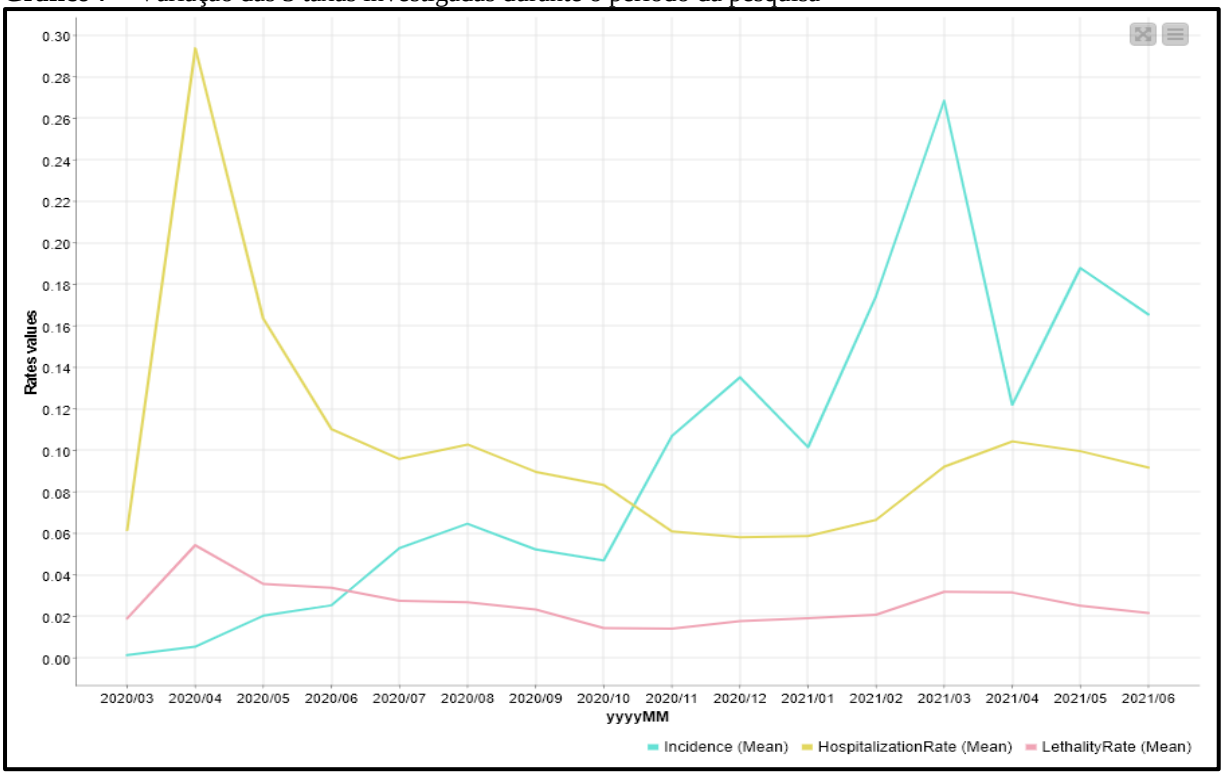
Para investigar se as supostas variáveis independentes atendem ao pressuposto da independência (ou de ausência de correlação) mútua, sua possível correlação foi investigada através do *node Linear Correlation*.

Tabela 5 – Correlação entre Incidência e Taxas de Hospitalização e Letalidade

Row ID	S First column name	S Second colu...	D Correlation value	D p value
Row0	Incidence	HospitalizationRate	0.0847828131790321	2.5575097595265106E-12
Row1	Incidence	LethalityRate	0.3011214424148037	0.0
Row2	HospitalizationRate	LethalityRate	0.5004155989186571	0.0

A análise dos valores acima, na investigação das correlações entre cada par de variáveis, suporta o uso das três taxas como variáveis efetivamente independentes entre si, o que autoriza a aplicação subsequente dos algoritmos que têm esta independência como um de seus pressupostos. O coeficiente de correlação linear de Pearson (r = 0,5) permite estimar que, no máximo, há somente uma associação mediana entre as taxas de Hospitalização e de Letalidade, porém nenhuma associação forte o suficiente para ser admitida como relevante para as análises subsequentes.

Gráfico 7 – Variação das 3 taxas investigadas durante o período da pesquisa



A independência entre as variáveis (supostamente) determinantes da flutuação nas métricas dos fluxos nos consultórios pôde ser inferida acompanhando-se suas variações temporais, como pela Análise Visual do Gráf. 7 e da Tab. 5, o que indica a validade do prosseguimento na investigação dos dados usando algoritmos que tenham como pressuposto uma independência de variáveis.

4.4 ETL E EDA DAS RESPOSTAS À *SURVEY*

A *survey* foi aplicada e esteve aberta a respostas durante o período de 01/07/2021 a 29/10/2021. Foram contatados aproximadamente 5.000 dentistas e consultórios odontológicos privados, gestores de clínicas (individuais ou redes de clínicas) privadas. Estes contatos foram feitos por meio de e-mails, envio de mensagens pelo site dos estabelecimentos ou de suas páginas na internet, de suas páginas em redes (como o Facebook, comunicação por WhatsApp e Instagram), ou por meio de ligações telefônicas e posterior envio dos convites de participação na pesquisa para o meio eletrônico indicado pelo profissional com quem fora feito o contato, e pela divulgação do link para a *survey*, tal como já havia sido feito no website da ABO-RS.

Não obstante os esforços da equipe de pesquisa para a captação de mais participantes, a adesão à pesquisa foi extremamente baixa. Foram coletadas 33 respostas, sendo que, destas, apenas 22 foram consideradas válidas. O critério de seleção quanto à validade das respostas fora definido *a priori* como sendo válidas as respostas enviadas por profissionais que fizessem a gestão (total ou parcial) dos estabelecimentos, ou de carteiras próprias de pacientes.

Os participantes estavam distribuídos, por município, conforme mostrado na Tab. 6:

Tabela 6 – Número absoluto e percentual por município de CDs/gestores participantes

Row ID	MUNICIP	NrRespsByMunic	NrProfs	REG_COVID	%Respondents
Row9	SANTANA DA BOA VISTA	1	12	PELOTAS - R21	8.33
Row0	ARROIO DOS RATOS	1	14	GUAIBA - R09	7.14
Row4	GAURAMA	1	15	ERECHIM - R16	6.67
Row7	RODEIO BONITO	1	17	PALMEIRA DAS MISSOES - R15 R20	5.88
Row1	CANGUCU	1	46	PELOTAS - R21	2.17
Row3	ENCANTADO	1	75	LAJEADO - R29 R30	1.33
Row10	TAPEJARA	1	99	PASSO FUNDO - R17 R18 R19	1.01
Row11	VIAMAO	1	126	PORTO ALEGRE - R10	0.79
Row5	PELOTAS	2	991	PELOTAS - R21	0.2
Row6	PORTO ALEGRE	10	6044	PORTO ALEGRE - R10	0.17
Row2	CANOAS	1	627	CANOAS - R08	0.16
Row8	SANTA MARIA	1	1184	SANTA MARIA - R01 R02	0.08

O processamento anterior (dos dados da COVID-19 no RS) havia gerado uma massa inicial de 6.795 linhas (de municípios-mês com registros de casos de COVID-19). Após o ETL das respostas à *survey*, e fazendo-se o cruzamento (uma vez mais, usando o *node Joiner*) entre os dados da *survey* e o ETL anterior, foi gerada uma redução na massa de dados para 182 linhas, restringindo-se os registros aos casos com respostas (à *survey)* nos mesmos meses. E foi a esta nova massa de dados que se procedeu à aplicação dos algoritmos subsequentes.

Assim, de um total de 24.644 CDs e EPAOs com registro no CRO-RS, apenas 0,09% (ou menos de 1 milésimo deste total) forneceram respostas válidas à *survey*. E, comparando as 22 respostas com o total de 9.250 profissionais registrados nestes mesmos 12 municípios, chegou-se a um valor de 0,24% de respondentes (Tabs. 7(a)-(b)).

Tabelas 7(a)-(b) – Totais e % de profissionais no RS e nos municípios dos participantes

Row ID	UF	Sum(SUM)	...	%CDsRS-RespsSurvey
Row0	RS	24644	...	0.08927122220418764

(a) % profissionais do RS que responderam à *survey*

ow ID	UF	Sum(NrProfs)	%CDsRS-RespsSurvey
Row0	RS	9250	0.23783783783783785

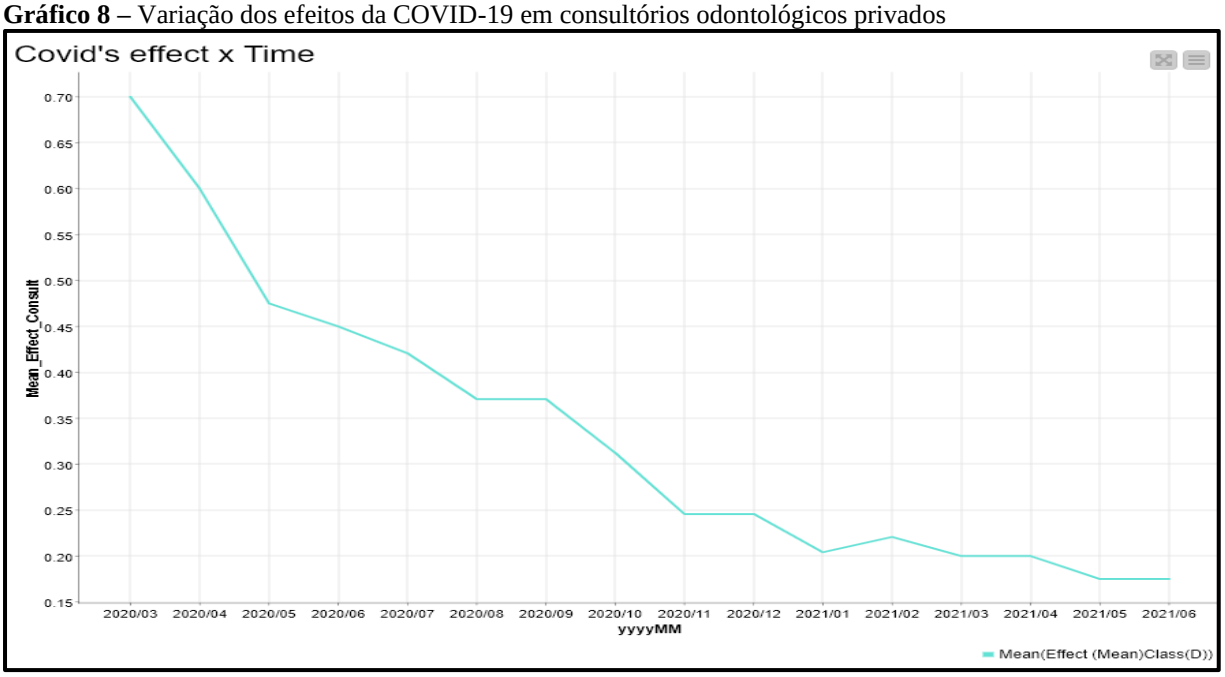
(b) % profissionais dos municípios que responderam

As respostas à *survey* – agrupadas por mês e tomadas por sua média – permitiram constatar uma redução constante no impacto da pandemia sobre a métrica (tomada como *proxy* (ou representativa das demais)), para as variações nas consultas efetivadas (Tab. 8 e Graf. 8).

Tabela 8 – Médias mensais da métrica das taxas e das consultas efetivadas

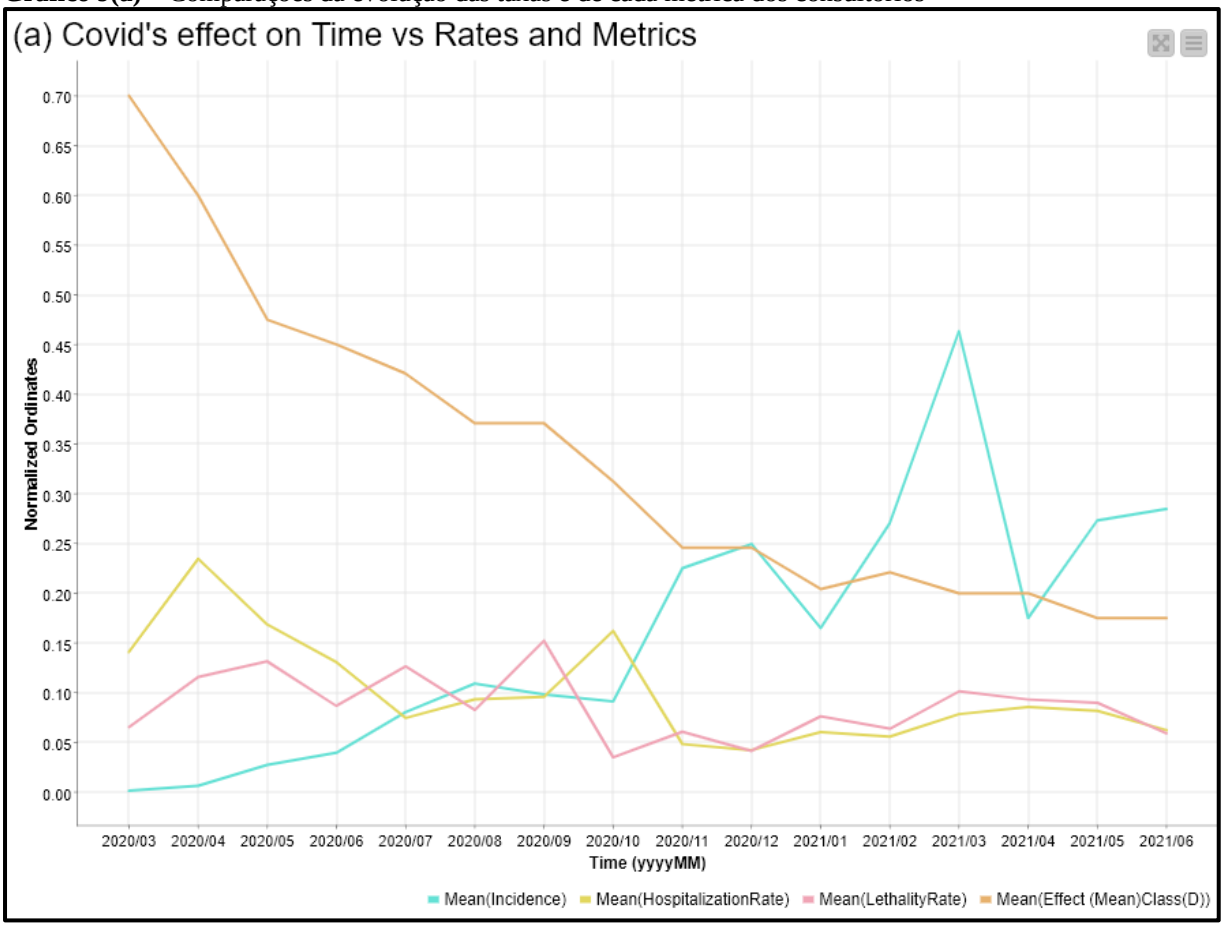
Row ID	S yyyy/MM	D Mean(Incidence)	D Mean(HospitalizationRate)	D Mean(LethalityRate)	D Mean(Effect (Mean)Class(D))
Row0	2020/03	0.001	0.141	0.065	0.7
Row1	2020/04	0.006	0.235	0.116	0.6
Row2	2020/05	0.027	0.168	0.131	0.475
Row3	2020/06	0.04	0.131	0.087	0.45
Row4	2020/07	0.08	0.075	0.127	0.421
Row5	2020/08	0.109	0.093	0.083	0.371
Row6	2020/09	0.098	0.096	0.152	0.371
Row7	2020/10	0.091	0.162	0.035	0.312
Row8	2020/11	0.225	0.048	0.061	0.246
Row9	2020/12	0.249	0.042	0.042	0.246
Row10	2021/01	0.165	0.06	0.076	0.204
Row11	2021/02	0.27	0.056	0.064	0.221
Row12	2021/03	0.463	0.078	0.101	0.2
Row13	2021/04	0.175	0.086	0.093	0.2
Row14	2021/05	0.273	0.082	0.09	0.175
Row15	2021/06	0.285	0.062	0.059	0.175

Gráfico 8 – Variação dos efeitos da COVID-19 em consultórios odontológicos privados



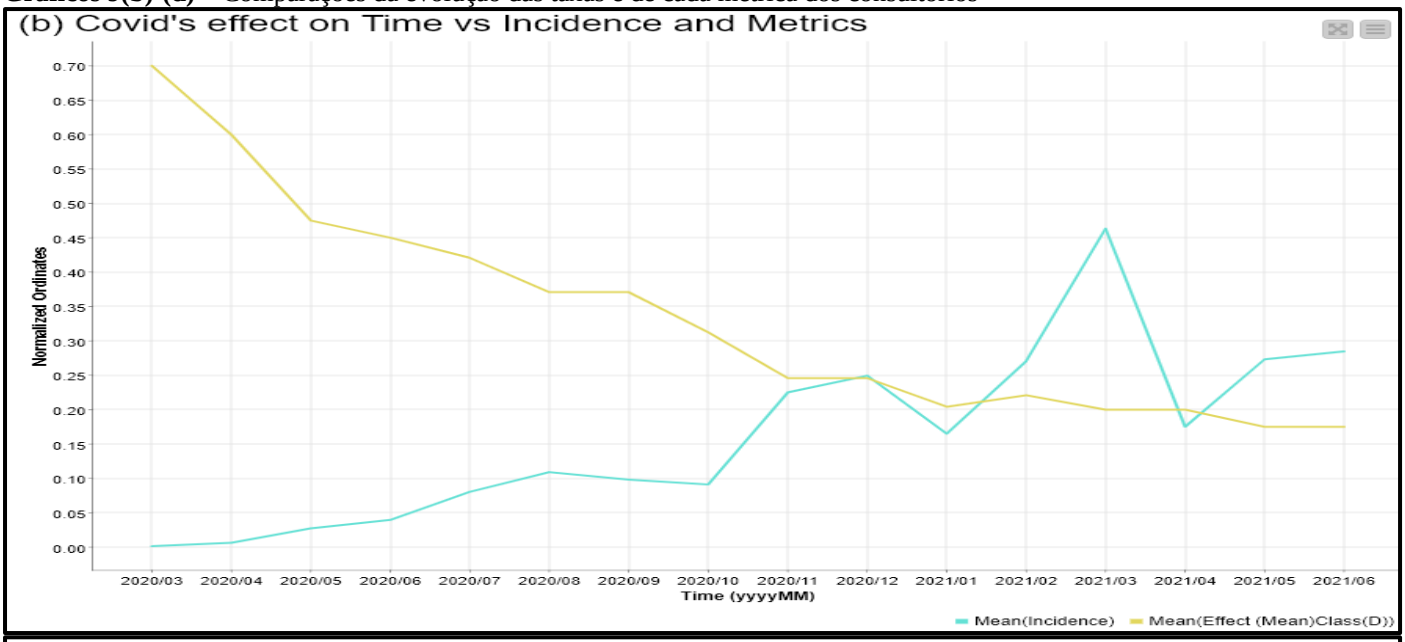
A análise visual do Graf. 9(a) e dos Gráfs. 9(b)-(d) mostra uma comparação entre a evolução temporal desta variável *proxy* e as das taxas usadas para avaliar a disseminação, a severidade e a severidade extrema da COVID-19 no RS. Esta análise indica, em que pese a falta de representatividade das respostas coletadas, como pode – ou, no caso desta pesquisa, como não pôde – ser observada uma colinearidade entre as taxas e os fluxos nos consultórios.

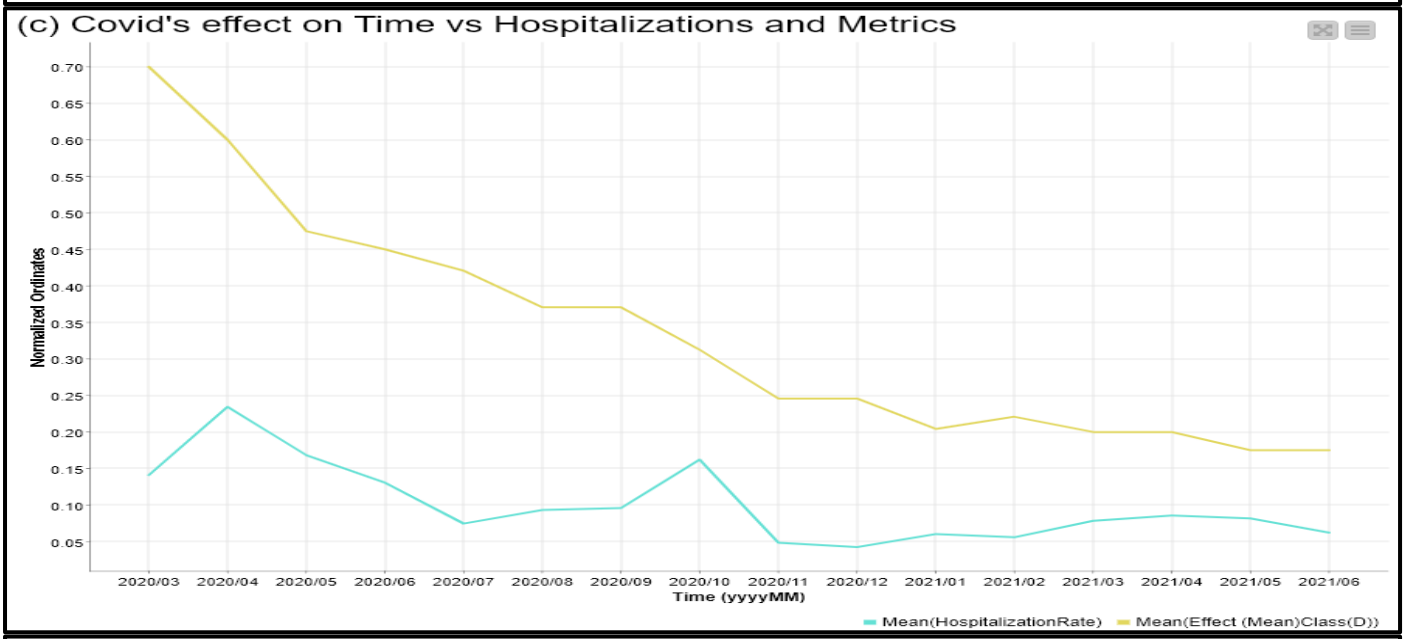
Gráfico 9(a) – Comparações da evolução das taxas e de cada métrica dos consultórios

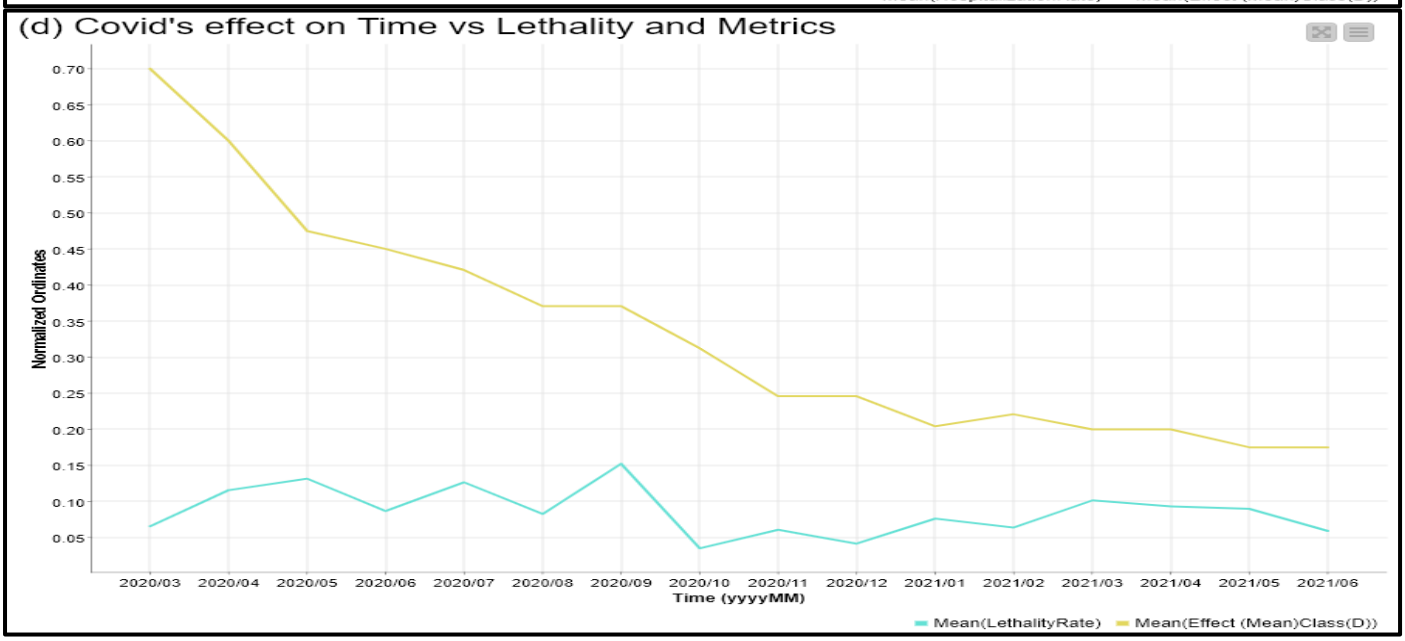


Com a finalidade de facilitar a análise visual destas curvas, foi feito o seu desdobramento, comparando individualmente cada uma das taxas (a da disseminação, a da severidade ou a da severidade extrema) com a métrica dos fluxos, nos Gráfs. 9(b)-(d). Esta análise visual segmentada usou o pareamento para simplificar a identificação de cada possível associação entre as alterações nos valores das variáveis que representam possíveis fatores determinantes (*i.e.*, as taxas) e nos valores da (suposta) variável dependente (*i.e.*, a métrica dos fluxos os consultórios), com a intenção de corroborar os resultados deste estudo. Foi considerada a inclusão de abordagens e técnicas que pudessem subsidiar ou evidenciar as associações buscadas. O mesmo vale para o correspondente método de análise, para que este possa ser adaptado e empregado na análise de questões de pesquisa similares.

Gráficos 9(b)-(d) – Comparações da evolução das taxas e de cada métrica dos consultórios







A Análise Visual das curvas acima parece sugerir que, na medida em que transcorreu o tempo de contato da sociedade com a COVID-19, o ritmo das frequências aos consultórios foi paulatinamente retornando a um fluxo mais próximo do anterior à pandemia, e que, das três variáveis investigadas (as citadas "taxas"), a única que apresentou uma redução foi a das hospitalizações, o que desperta o interesse investigativo para possíveis associações entre a redução na severidade dos casos e o retorno do fluxo nas consultas à normalidade, mais do que no ritmo de disseminação (que cresceu, no período), ou o da severidade extrema (que se manteve em patamares muito próximos da estabilidade). Para investigar este tópico através da busca de uma correlação linear (de Pearson), foi usado o *node Linear Correlation* entre cada par de variáveis (Tab. 9).

Tabela 9 – Correlação entre variação nos fluxos de consultas e as taxas da COVID-19

Row ID	S First colu...	S Second column name	D Correlation value	D p value
Row0	Mean(Incidence)	Mean(HospitalizationRate)	-0.6805873831976...	0.003710201806442...
Row1	Mean(Incidence)	Mean(LethalityRate)	-0.2754637284646...	0.3017778673320046
Row2	Mean(Incidence)	Mean(Effect (Mean)Clas...	-0.8164876186157...	1.139497153839805E-4
Row3	Mean(Hospitali...	Mean(LethalityRate)	0.29053972202093	0.274985784170769
Row4	Mean(Hospitali...	Mean(Effect (Mean)Clas...	0.7410355946297141	0.001021935124343...
Row5	Mean(Lethality...	Mean(Effect (Mean)Clas...	0.30997504954311...	0.24265231336478976

Esta busca identificou duas associações lineares significativas (*i.e.*, no caso da presente pesquisa, com *p*-valor < 0,05), relativas aos efeitos da COVID-19 no mercado-alvo (o que é o objetivo central desta pesquisa): a) uma direta, de moderada a forte (r = 0,74), calculada entre a taxa de hospitalizações e a métrica para os fluxos nas consultas nos consultórios dos participantes; e b) outra inversa, e forte (r = - 0,82), calculada entre a Incidência e a mesma métrica dos fluxos.

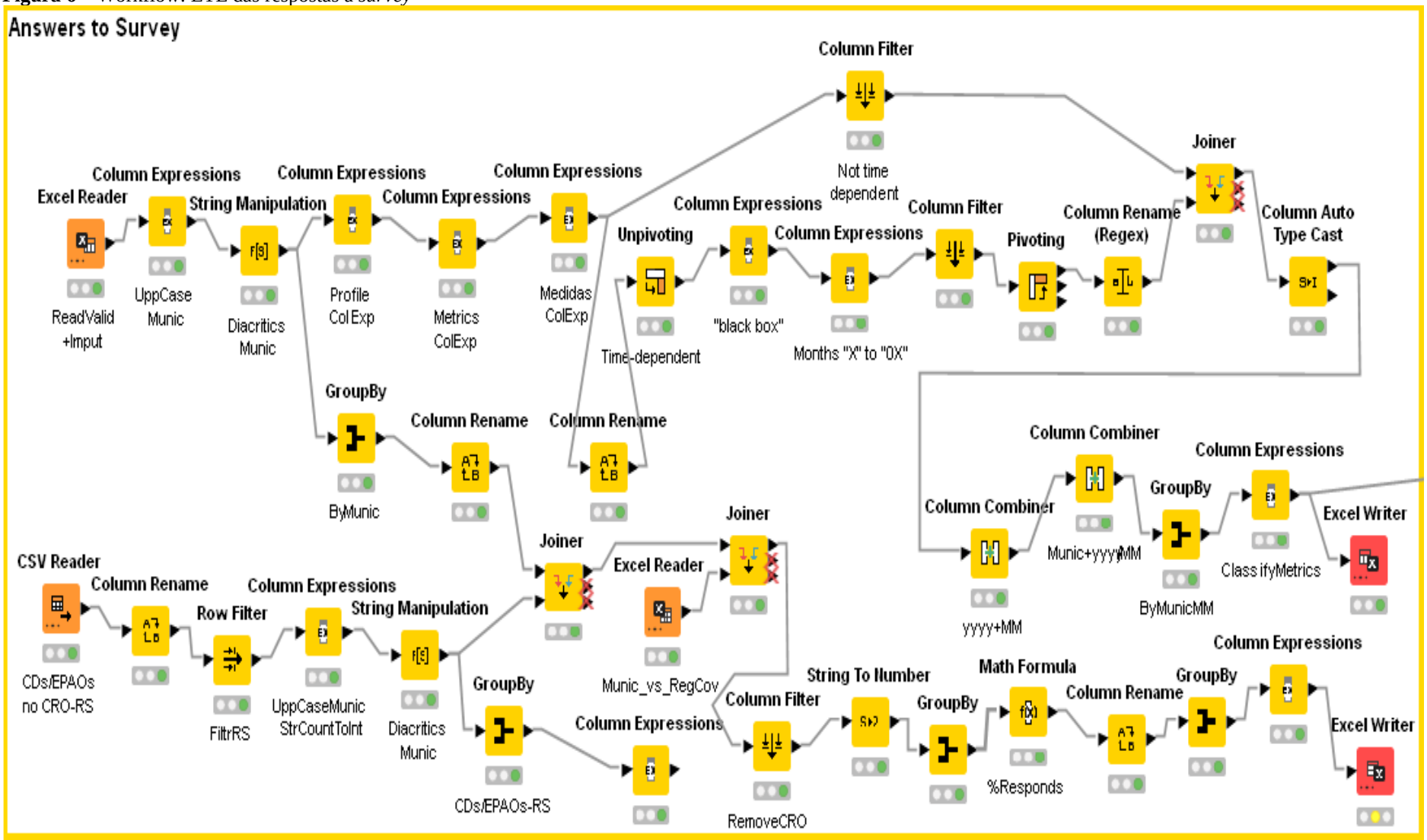
A sistematização das respostas à *survey* foi feita na Tab. 10, com as classes e as respectivas contagens e frequências para a variação nos fluxos dos consultórios.

Tabela 10 – Respostas coletadas na *survey* – contagem e frequência das classes

D Effect (Mean)Class(D)	I Count (Effect (Mean)Class(D))	D Relative Frequency (Effect (Mean)Class(D))
0.8	21	0.115
0.6	29	0.159
0.4	30	0.165
0.2	31	0.17
0.1	49	0.269
0.05	22	0.121

O trecho de workflow correspondente ao ETL das respostas à *survey* (sem o uso de *metanodes* ou de *components*) pode ser observado na Fig. 6, e também é interpretado, nos trechos correspondentes do texto referente ao ETL, a partir dela.

Figura 6 – Workflow: ETL das respostas à *survey*



4.5 TAREFAS DE ML NO KNIME

A partir desta seção foram aplicadas estratégias de ML, buscando fazer o KDD através dos algoritmos de ML selecionados e brevemente descritos nas seções correspondentes a cada um destes algoritmos.

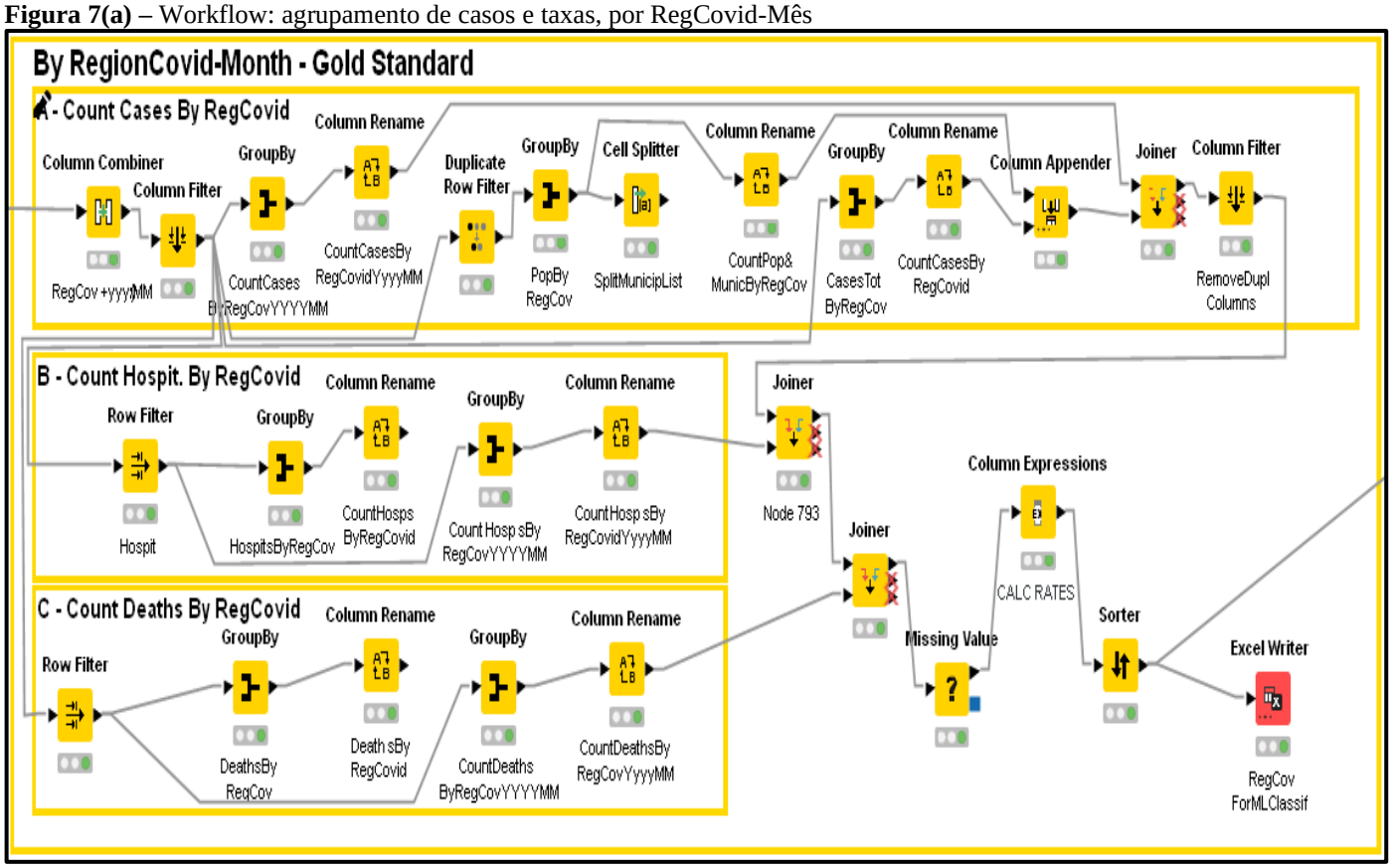
As tarefas desenvolvidas nesta primeira etapa, como preparação para as demais atividades de ML, envolveram as seguintes atividades:

a) agrupamento dos casos de, e hospitalizações e óbitos por COVID-19, conforme as "Regiões Covid" (o padrão adotado pela SES-RS) e os meses em que estes foram registrados, para caracterizar este estado dinâmico de cada Região Covid, ao longo do período pesquisado. O trecho do workflow correspondente está na Fig. 7(a). Este agrupamento representa o *gold standard* e foi usado para comparação com os resultados do modelo gerado na pesquisa;
b) agrupamento (ou clusterização) dos casos conforme os municípios e meses em que estes foram registrados. O trecho do workflow correspondente está na Fig. 7(b). Este agrupamento foi usado no desenvolvimento, aplicação e validação dos modelos construídos no decorrer da pesquisa. O objetivo do agrupamento dos casos por "município-mês" foi o de buscar métodos alternativos (à por proximidade geográfica) para tornar comparáveis diferentes municípios que tenham, em um dado mês, situações análogas quanto às métricas da COVID-19, para daí extrair informação relevante (KDD) para subsidiar melhores decisões frente aos efeitos do crescimento (ou arrefecimento) na forma como a pandemia atinge cada município.

As atividades de agrupamento (ou clusterização) dos casos conforme os municípios e meses de registro são representados pelo trecho de workflow que está na Fig. 8. Este agrupamento serviu para o desenvolvimento, aplicação e validação dos modelos construídos no decorrer da pesquisa. O objetivo do agrupamento dos casos por "município-mês" deve-se à intenção de buscar métodos alternativos (ao da clusterização por proximidade geográfica, que já é o padrão vigente na administração da saúde pública no RS). E foi definido que as tarefas de ML usassem as três taxas de trabalho para que, neste modelo, diferentes municípios fossem comparáveis, por ter, em um dado mês, situações análogas quanto às métricas da COVID-19. Pretendeu-se, a partir daí, extrair informação relevante (KDD) para subsidiar melhores decisões frente aos efeitos do crescimento (ou arrefecimento) na forma como a pandemia atinge diferentes municipalidades a cada momento (ou período) observado.

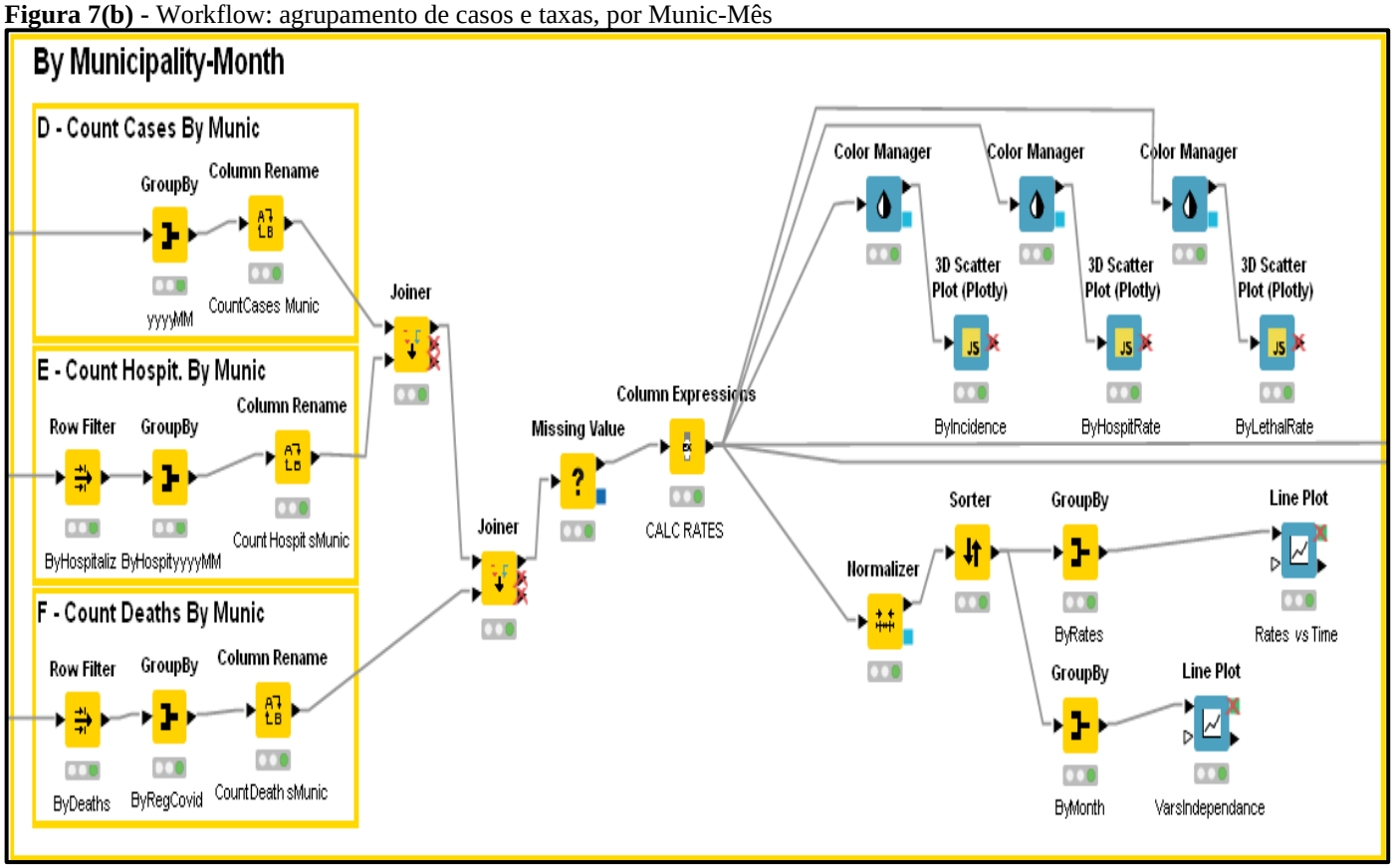
A Fig. 7(a) mostra o workflow correspondente ao agrupamento dos casos de, e das hospitalizações e óbitos por COVID-19, segundo a "RegCovid-Mês", para caracterizar este estado dinâmico de cada Região Covid, ao longo do período pesquisado. Analogamente, a Fig. 7(b) mostra o agrupamento por "Munic-Mês", para fins análogos relativos aos municípios.

Figura 7(a) – Workflow: agrupamento de casos e taxas, por RegCovid-Mês



c) cálculo das três citadas taxas municipais (e suas médias) – as quais servem para a comparação entre grupos de dados com diferentes magnitudes absolutas individuais – convertendo-os em pontos diretamente comparáveis entre si.

Figura 7(b) - Workflow: agrupamento de casos e taxas, por Munic-Mês



A partir deste trecho, são relatadas as seguintes atividades de ML feitas, com os respectivos resultados, visando atingir os objetivos da pesquisa aqui apresentada:

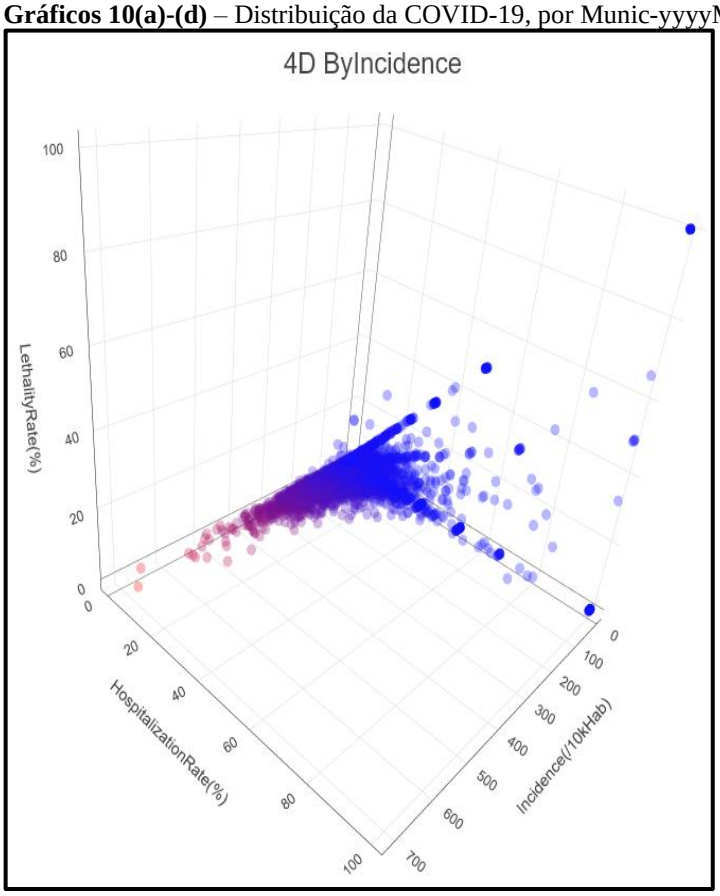
d) análise da qualidade dos *clusters* gerados (ou performance deste tipo de agrupamento de acordo com as variáveis selecionadas) através da aplicação de seus Coeficientes de Silhueta Médios (CSMs), com a otimização do número *k* de *clusters* tão coesos e distintos entre si quanto possível;
e) seleção dos dados de (d), acima, e seu agrupamento conforme o número otimizado de *k clusters*;
f) ETL dos dados obtidos dos questionários para um formato diretamente comparável aos da massa geral de dados dos pacientes de COVID-19 no RS;
g) filtragem dos dados de (e), acima, restringindo-os somente aos municípios para os quais houve respostas válidas ao questionário, em (f), acima → esta tarefa visou agregar os dados epidemiológicos municipais da COVID-19 às métricas dos consultórios para o posterior pareamento e análises com recursos de DS, o que foi feito com o *node Joiner*;
h) aplicação de algoritmos de regressão e de classificação (ou de predição de valores para as variáveis respectivamente numéricas ou categóricas) dos efeitos econômicos da COVID-19 nos membros dos diferentes *clusters* → esta tarefa visou a descoberta de valores mais prováveis para alterações nas métricas dos consultórios como função das taxas municipais de disseminação e gravidade da pandemia em um dado município e em um mês específico.

Os dados foram agrupados conforme as três taxas já descritas e calculadas diretamente com os *nodes* do Knime, para cada município e em cada mês. Assim, cada "ponto" ou observação possui três coordenadas em um espaço euclidiano tridimensional (3D), o das três ordenadas espaciais (das taxas individuais) para cada município e cada mês. Dados os 497 municípios do RS e os 16 meses da pesquisa, poder-se-ia obter um máximo de 7.952 observações. Porém, considerando-se que nem todos os municípios tiveram casos registrados de COVID-19 em todos os meses, o número real computado de observações ficou respectivamente em 6.795 observações (ou pontos) de registros, 5.092 hospitalizações e 3.478 óbitos, a serem agrupados em *clusters* conforme sua situação municipal em um dado mês.

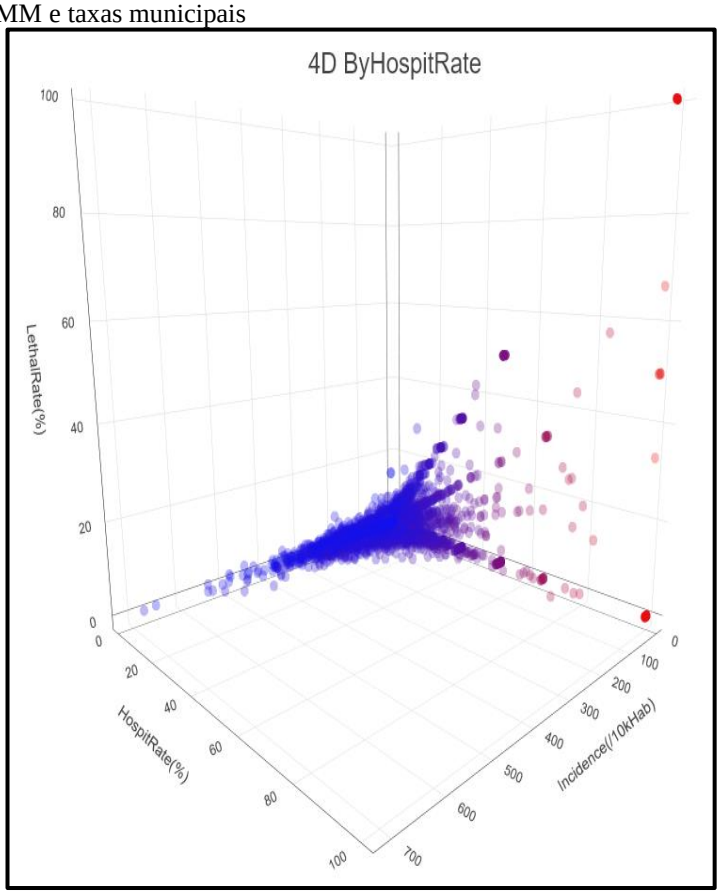
Para o desenvolvimento de uma aplicação de DS na busca de solução para um problema de pesquisa, os dados coletados podem ser representados segundo um número variável de dimensões. No presente trabalho, foram adotadas as três citadas taxas (ou dimensões), usadas nas análises descritas a seguir, e incorporam uma quarta dimensão, a do tempo (em meses).

Os Gráfs. 10(a)-(d) mostram a dispersão desta distribuição, por gradiente de cor, usando cada taxa em um eixo ortogonal, para representar o gradiente das diferenças relativas de distribuição segundo as **incidências** e taxas locais mensais de **hospitalização** e de **letalidade** por COVID-19 no RS no período da pesquisa, em que cada ponto representa o subconjunto de casos mensais de um município. Também mostram um possível reordenamento em subgrupos.

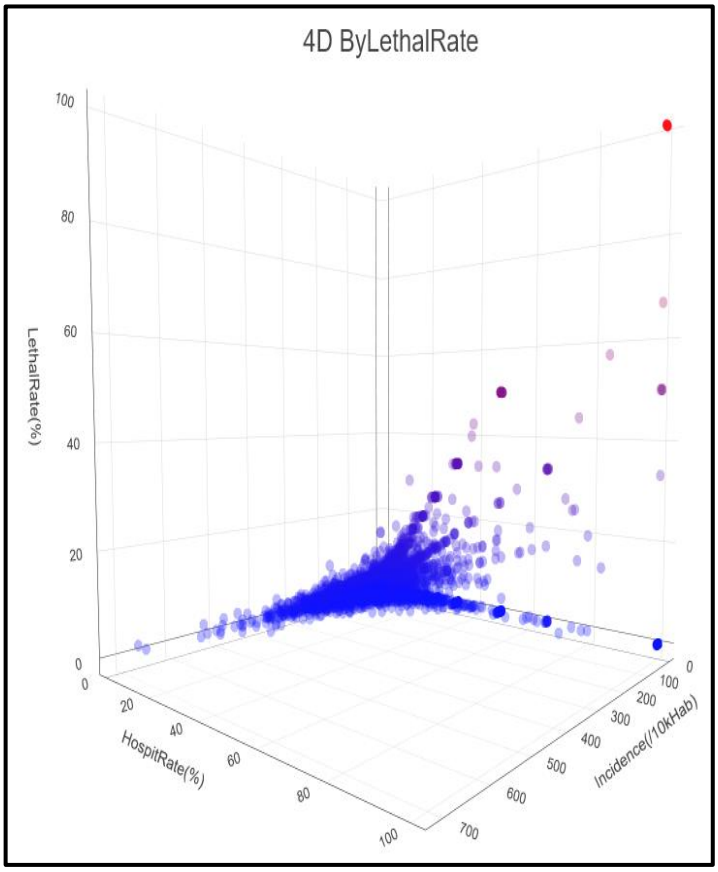
Gráficos 10(a)-(d) – Distribuição da COVID-19, por Munic-yyyyMM e taxas municipais



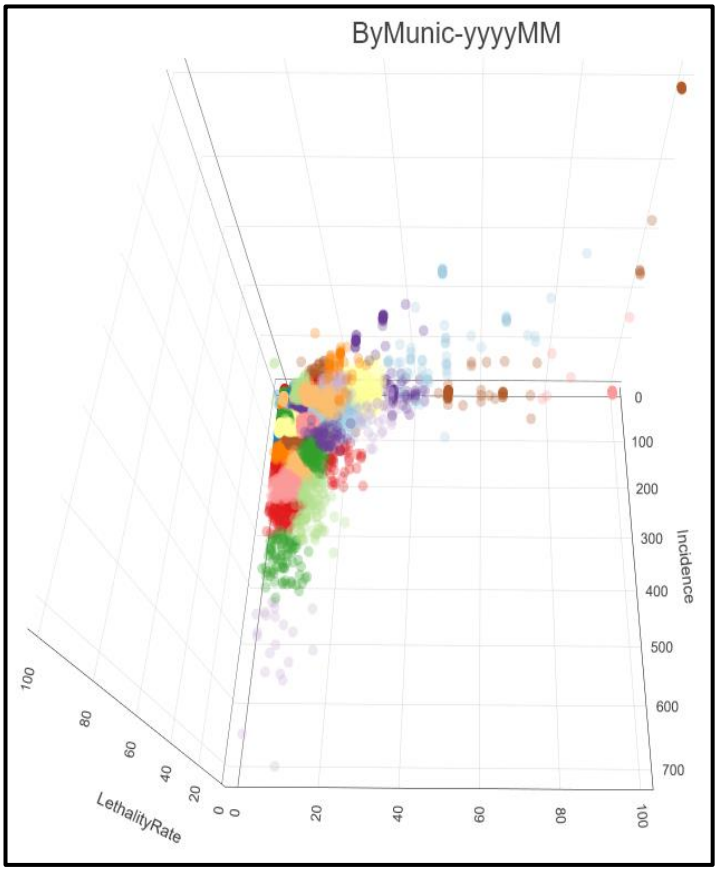
(a) por incidência



(b) por Taxa de Hospitalização



(c) por Taxa de Letalidade



(d) para um n° grande (= 50) de *clusters* para as taxas

Visualizações como as acima (*i.e.*, em (a)-(c), em gradientes por variáveis contínuas, ou em (d), com muitos pequenos agrupamentos bem mais coesos) não deixam muito distintas as eventuais associações ou similaridades entre diferentes subgrupos de situações mensais locais. No entanto, parecem sugerir que ao menos parte destes pontos tenham mais características em comum com outros, e que possam ser reagrupados conforme estas três dimensões. E isto pode indicar que o procedimento planejado – o de testar a aplicação de diferentes algoritmos para identificar qual(ais) deles pode(m) melhor retratar como estratificar uma massa de dados específica – pode ser relevante para caracterizar adequadamente como sub-agrupar estes dados.

Portanto, passa-se agora à descrição dos resultados obtidos com o algoritmo selecionado para a clusterização dos dados.

4.5.1 Da clusterização dos dados com o algoritmo k-Means

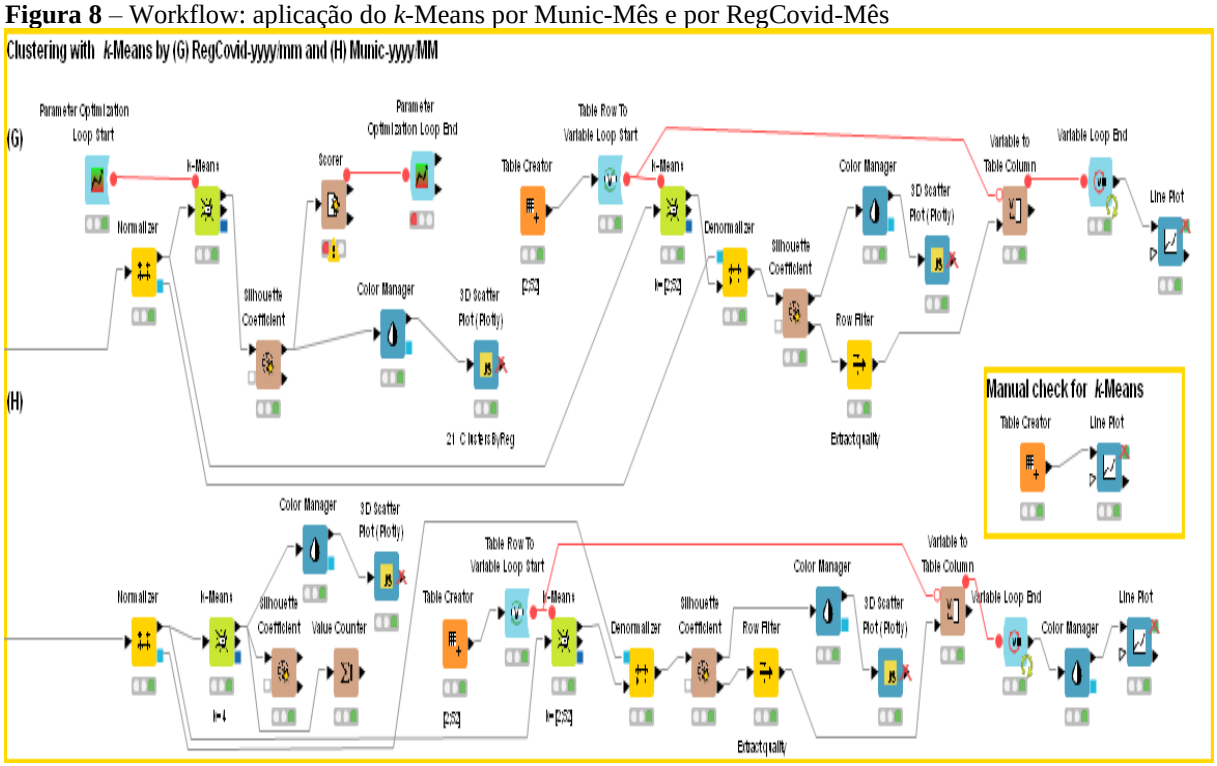
Antes da clusterização propriamente dita, os dados brutos foram pré-processados e reagrupados no segmento de workflow precedente (o ETL), que chegou a um número total (*NCases*) de 1.324.589 casos devidos à COVID-19, parte deles evoluindo para hospitalizações, e destes, uma parte menor para o êxito fatal.

Procedeu-se então a um procedimento de ML não supervisionado, o do *clustering*, e para esta tarefa foi selecionado o algoritmo *k-Means*. Este algoritmo foi selecionado, dentre diversos possíveis outros, por diferentes razões, a começar por ser marcadamente o mais popular dentre os pesquisadores, e também por ser o mais facilmente comparável aos de outras pesquisas, o mais abrangente e includente, entre outros argumentos. Nesta pesquisa, foi feito o reagrupamento dos dados em dois *subsets* nos ramos abaixo:

(a) dos dados brutos, conforme suas taxas locais mensais;
(b) dos dados trabalhados nos seus ramos (D), (E) e (F) – os do agrupamento por Municipality-month, analogamente ao feito por RegCovid-month.

A Fig. 8 mostra esquematicamente estas duas linhas (ou ramos) separados do workflow no k-Means. Para avaliar a qualidade do processo de clusterização, foi adotado o Elbow Method, para a definição do número mais otimizado de *k clusters*, pelos seus CSMs.

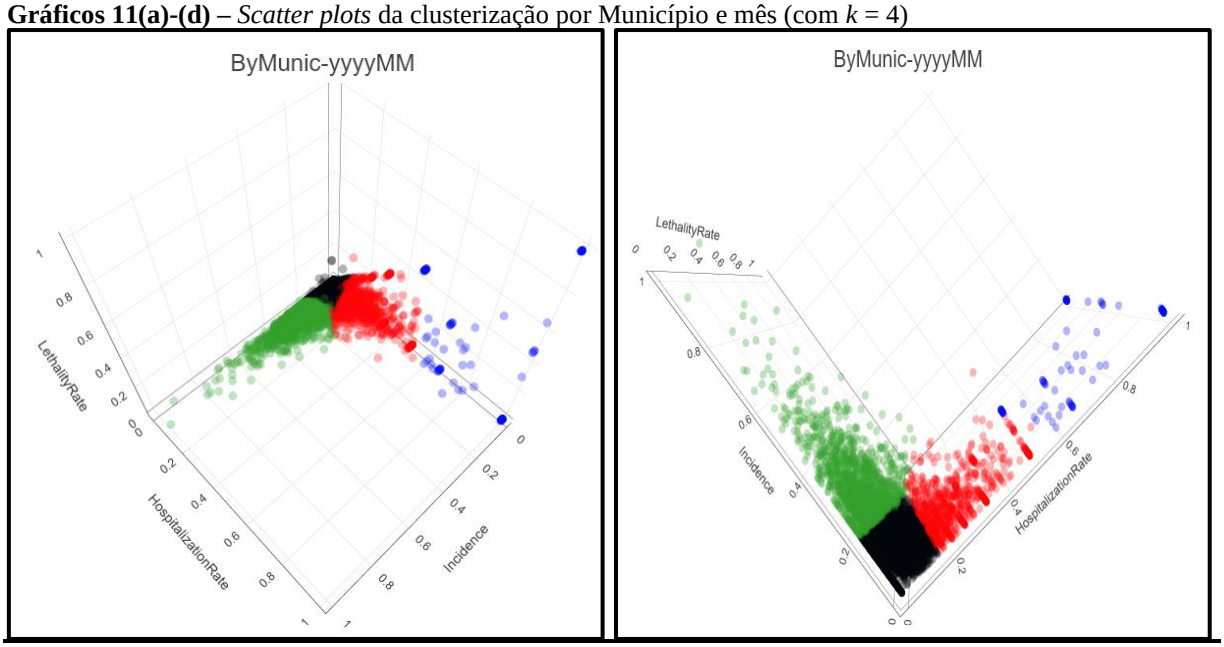
Figura 8 – Workflow: aplicação do *k*-Means por Munic-Mês e por RegCovid-Mês



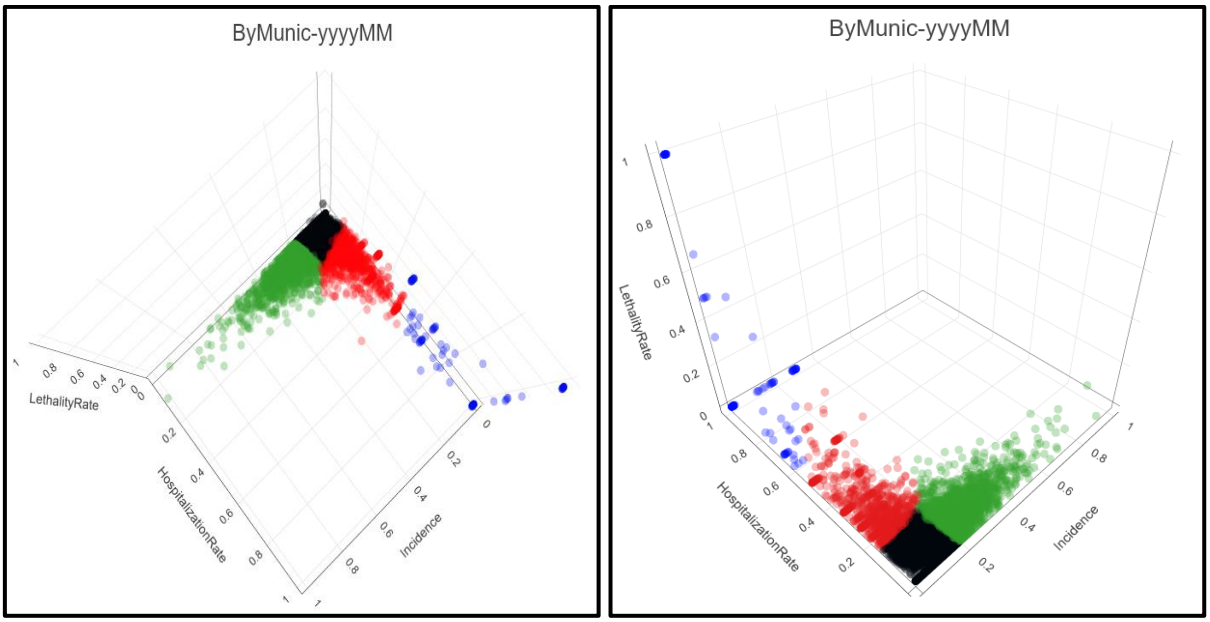
Para estas clusterizações, destaca-se a partir daqui os resultados obtidos com a principal linha de trabalho do modelo, que foi a do agrupamento por município-mês (H).

A primeira foi a aplicação de uma rotina de otimização do número *k* de *clusters* através da avaliação pelo CSM (em um *loop* de otimização de parâmetros), o que resultou na escolha do agrupamento (em observações município-Mês) em 4 *clusters*, conforme mostrado abaixo, em quatro diferentes perspectivas (POV) do Gráf. 11(a)-(d).

Gráficos 11(a)-(d) – *Scatter plots* da clusterização por Município e mês (com *k* = 4)



(a) Isometric view (b) bottom-up view



(c) Top-down view (d) rear view

A distribuição de pontos em cada um destes 4 *subsets* pode ser observada na Tab. 11.

Tabela 11 – Distribuições das taxas de COVID-19 por *cluster*

Cluster_#	Observations # (Total)	Observations (%)	Color*
2	4416	64.99	**Black**
0	1394	20.51	**Green**
1	851	12.52	**Red**
3	134	1.97	**Blue**

* A cor refere-se ao *scatter plot* (gerado no Knime) para o k-otimizado (Gráf. 11(a)-(d)).

a) um deles (o *Cluster*_2) agrupou municípios em meses que apresentaram todas as três taxas baixas. Reuniu 64,99% de todos os pontos (ou 4416 dos 6795);
b) o segundo (por ordem de número de componentes) foi o *Cluster*_0, que reuniu 20% dos pontos (1394), que foram os municípios-mês com as incidências mais altas;
c) o terceiro foi o *Cluster*_1, com 851 pontos (ou 12,52% do total), que agregou os municípios-mês em que as incidências ficaram nos 20% inferiores, porém com as taxas de hospitalização e de letalidade nos 50% inferiores; e
d) o quarto foi o *Cluster*_3, com apenas 134 pontos (ou quase 2% (1,97%) do total), reunindo as taxas mais elevadas tanto de hospitalização (todas as instâncias acima de 55%) quanto de letalidade (todas as instâncias acima de 43%), porém concentradas abaixo de 10% das incidências.

A fase subsequente do workflow incluiu a geração automatizada de um gráfico para a determinação do melhor número *k* de *clusters*, que foi verificada manualmente, conforme mostrado no Gráf. 9(a), com o primeiro mínimo local em $k = 4$; e no Gráf. 9(b), com um máximo local (e provavelmente também o máximo global) em $k = 4$.

Gráfico 12(a) – Curva *k vs* MSC – *Elbow Method* na otimização do número de *clusters*

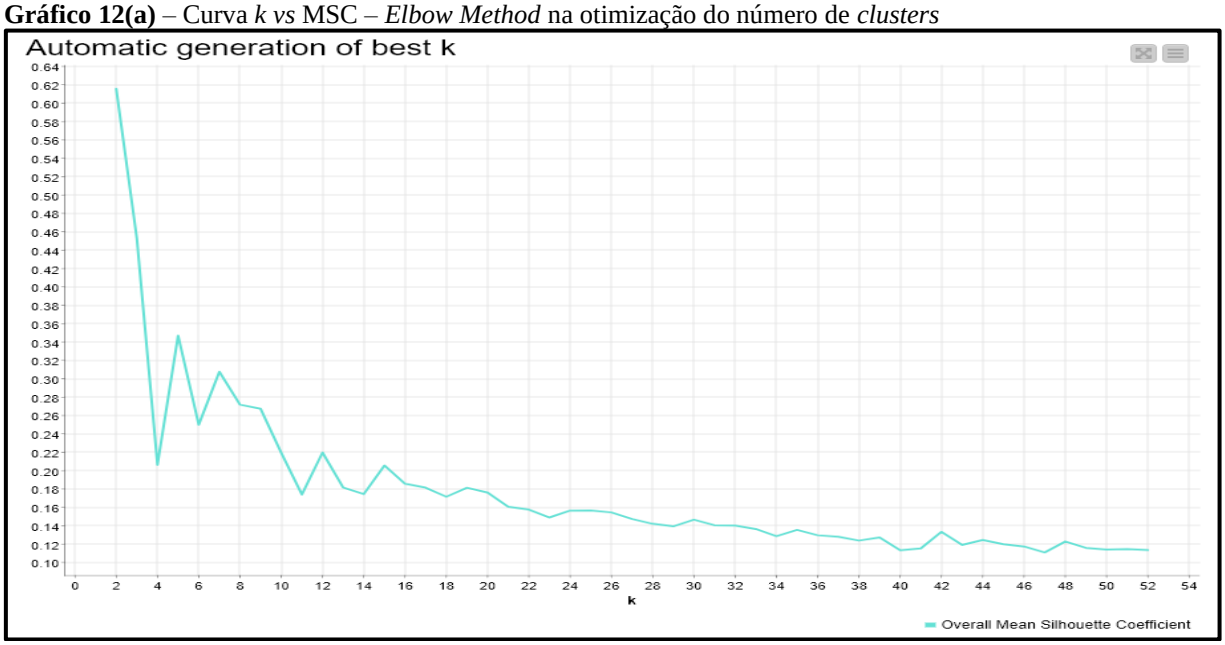
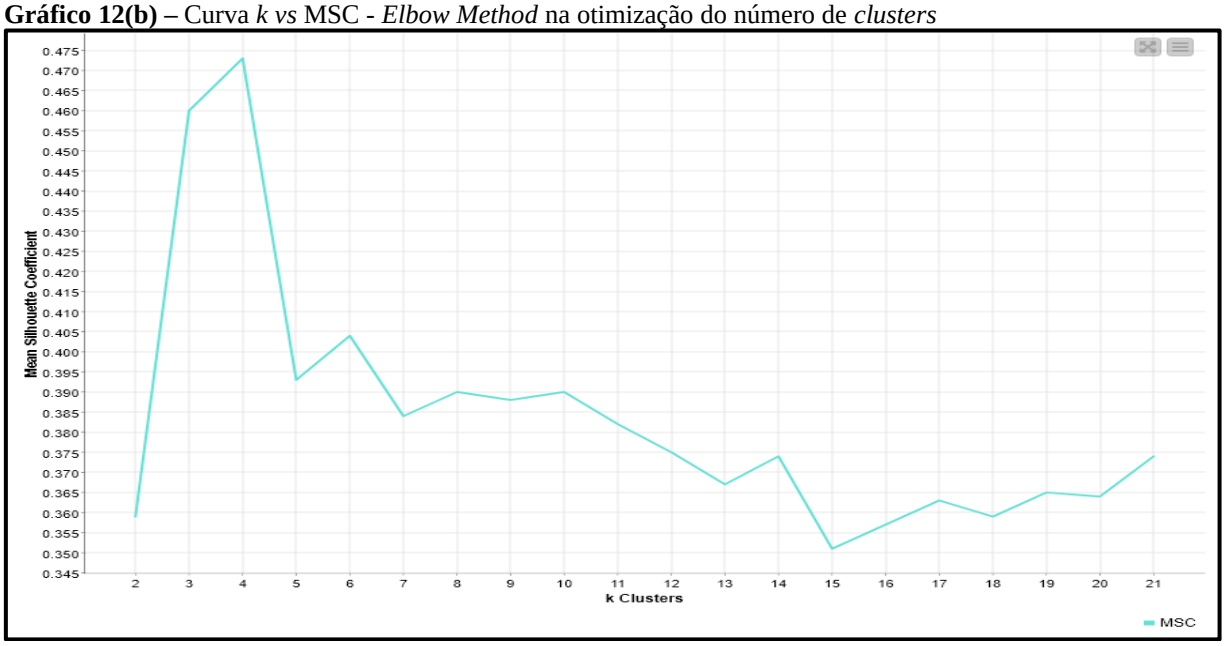


Gráfico 12(b) – Curva *k vs* MSC - *Elbow Method* na otimização do número de *clusters*



Estes gráficos representam o maior nível possível de coesividade *intracluster* e de distinção *intercluster* de acordo com a avaliação pelos seus CSMs. Na Tab. 12(a)-(b) são mostrados os centroides de cada *cluster*, nas dimensões selecionadas, com e sem normalização.

Tabelas 12(a)-(b) – Coordenadas (com e sem normalização) dos centroides dos *clusters*

Row ID	Incidence	HospitalizationRate	LethalityRate
cluster_0	0.283	0.071	0.021
cluster_1	0.07	0.046	0.013
cluster_2	0.008	0.856	0.222
cluster_3	0.046	0.256	0.06

(a) com normalização

Incidence	HospitalizationRate	LethalityRate
200.425	7.144	2.135
49.841	4.622	1.315
6.056	85.596	22.164
32.34	25.63	5.972

(b) sem normalização

E, por meio do emprego de um dos *nodes* agrupados dentre as funções de Estatística, como o próprio Statistics ou o Value Counter, obtém-se a contagem de quantos municípios-mês foram atribuídos a cada um destes quatro *clusters*, como mostrado na Tab. 13.

Tabela 13 – Contagem de Munic-Mês / *cluster*

Row ID	count
cluster_0	1394
cluster_1	851
cluster_2	4416
cluster_3	134

E, na Tab. 14 são mostrados os CSMs de cada *cluster* e o geral.

Tabela 14 – CSMs geral e dos *clusters*

Row ID	Mean Silhouette Coefficient
cluster_2	0.543
cluster_0	0.383
cluster_1	0.276
cluster_3	0.351
Overall	0.473

Uma vez aplicado o algoritmo não supervisionado do *k-Means* para a clusterização, e tendo já sido descoberto – pelo *Elbow Method*, usando os CSMs – o valor otimizado para o número *k* de subgrupos tão coesos e tão distintos entre si quanto possível, passa-se para os demais algoritmos de predição, sejam estes os "de regressão" (tipicamente aplicado a variáveis contínuas) ou os "de classificação" (tipicamente aplicado a variáveis categóricas), sabendo-se já as "classes[8]" (ou valores de resposta das variáveis preditas) como função das preditoras.

4.5.2 Resultados das predições pelas Regressões Linear, Polinomial e Logística

Os algoritmos de Regressão Linear (*LinReg*) e Polinomial (*PolyReg*) buscam as curvas, com seus diferentes graus (*i.e.*, respectivamente polinômios de grau 1 e de grau ≥2) que passem o mais próximo possível da maioria dos pontos da massa de dados de teste, após o treinamento do algoritmo, tal como apresentado no trecho de workflow da Fig. 9. E os gráficos das curvas ROC (Apêndice D) para as diferentes classes de resultados, obtidas com a aplicação destes

[8] Toma-se o conceito de "classe" como cada um dos possíveis conjuntos (ou intervalos) em que possam ser agrupados (ou "classificados") os valores das variáveis dependentes (*e.g.*, as respostas já coletadas ou futuros dados a serem preditos pelos algoritmos) a partir do treinamento com os dados já conhecidos e trabalhados por este aplicativo.

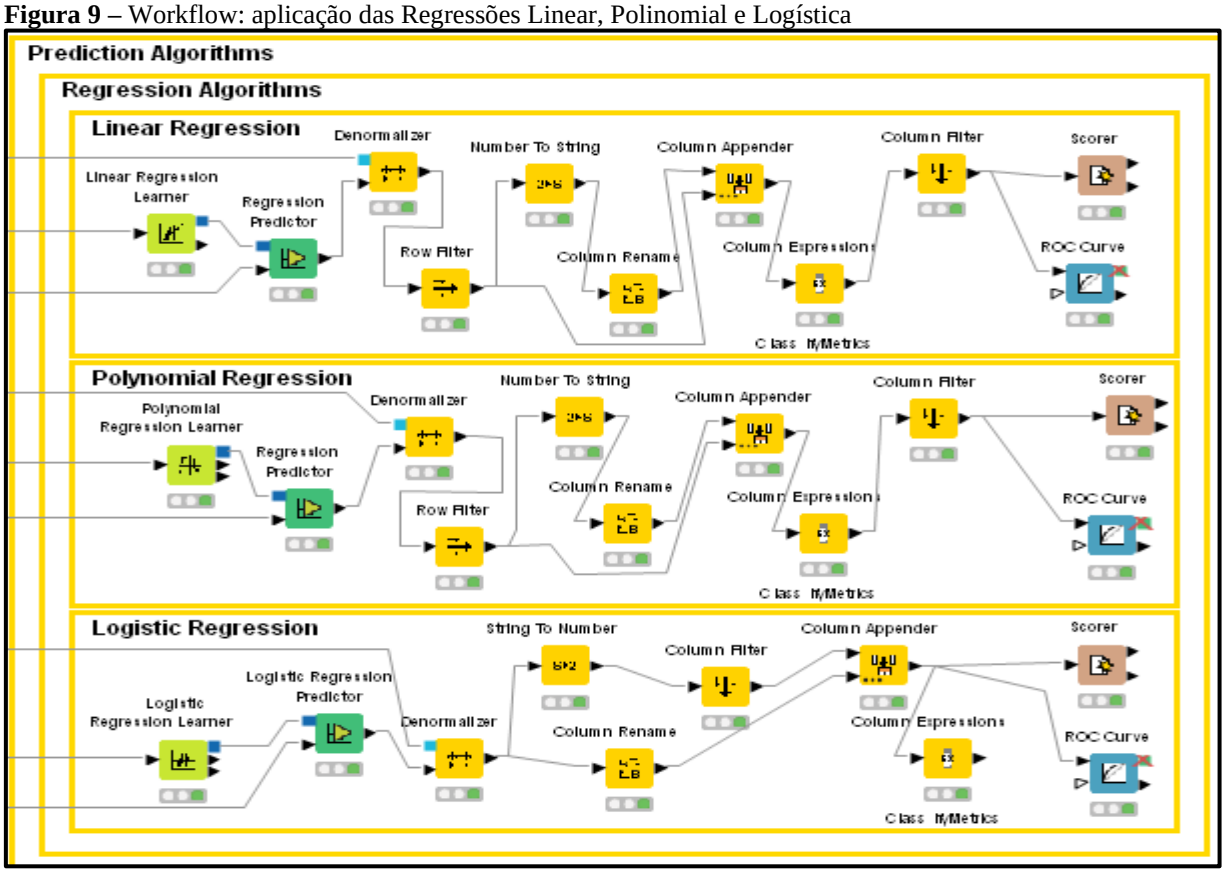
algoritmos para a massa de dados resultante da combinação entre: a) os dados do panorama geral da COVID-19 no RS, com dados de 6795 entradas Municípios-Mês; e b) restritos às 22 respostas dadas à *survey*, obtidas dos 12 municípios com respostas válidas. Este cruzamento (feito em um *node Joiner*) gerou 182 entradas Municípios-Mês, trabalhadas nos algoritmos de ML do restante dos procedimentos da pesquisa aqui apresentada.

As análises desta seção iniciam com a apresentação dos trechos do workflow referente às Regressões Linear (*LinReg*), Polinomial (*PolyReg*) e Logística (*LogReg*). Segue-se com a análise para cada regressão separadamente, incluindo sua Matriz de Confusão (MC) e uma sequência de gráficos das curvas ROC, respectivamente para as sucessivas classes em que foram agrupadas as possíveis respostas (estando estas últimas no Apêndice D), com a avaliação das AUCs sintetizadas na Tabela 22. Compara-se as taxas da COVID-19 com a métrica real (apurada pela *survey*) da variação nos fluxos dos consultórios, e com a predição para esta métrica em cada algoritmo segundo o modelo, destacando em negrito as AUC da métrica de melhor ajuste (arbitradas como representando $\geq 2/3$ de predições certas).

Apresenta-se aqui os resultados e a avaliação das performances dos algoritmos de regressão, conforme a análise das respectivas MCs. A Tabela 22 descreve as análises, e sintetiza as AUCs das curvas ROC para cada regressão e classificação.

A Fig. 9 mostra os trechos de workflow correspondentes à aplicação dos 5 algoritmos de regressão que foram selecionados para teste, execução e comparação nesta pesquisa.

Figura 9 – Workflow: aplicação das Regressões Linear, Polinomial e Logística



Após a apresentação dos trechos de workflow relativos ao design e à configuração dos cindo algoritmos de regressão selecionados, o primeiro deles, o da Regressão Linear (*LinReg*) foi executado, permitindo a avaliação de sua performance por meio do nível de acurácia apresentado em sua Matriz de Confusão (Tab. 15).

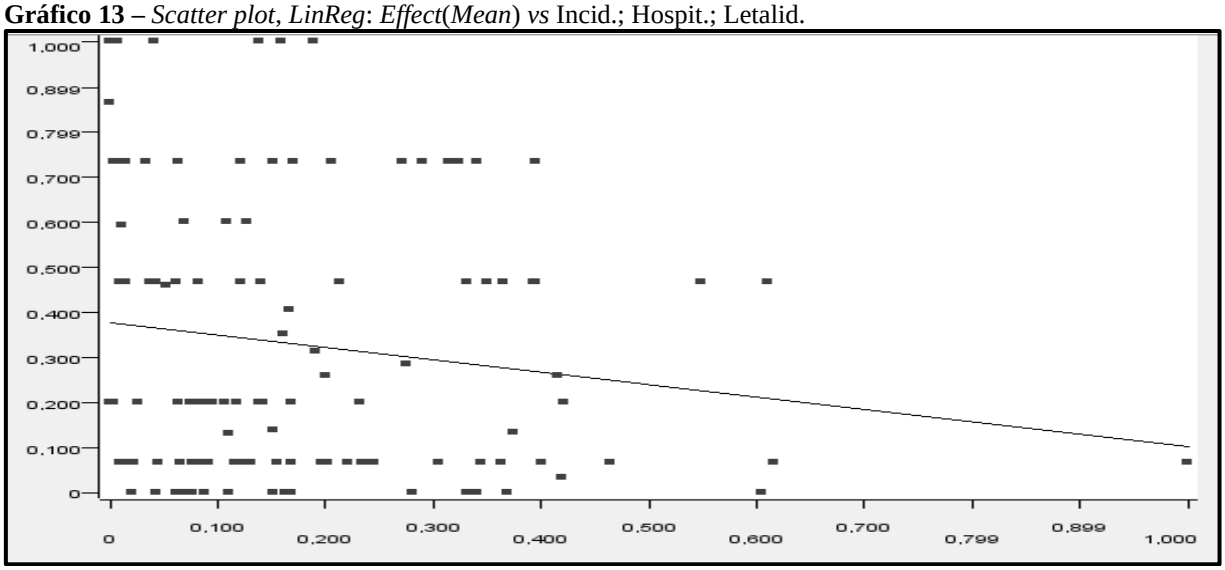
Tabela 15 – Matriz de Confusão da *LinReg*

rounded(Efet (Mean) \ round(Predict(round((efet)) (#1)	0.8	0.6	0.4	0.2	0.1	0.05
0.8	1	2	5	0	0	0
0.6	1	2	7	0	0	0
0.4	1	0	6	0	0	0
0.2	0	1	10	0	0	0
0.1	0	1	12	0	0	0
0.05	0	0	5	0	0	0
Correct classified: 9			Wrong classified: 45			
Accuracy: 16,667%			Error: 83,333%			

A análise da MC, das acurácias apresentadas pelo modelo, e dos mostra que, para a maior parte das classes de resposta, das observações e das variáveis – exceto em algumas delas, e geralmente para apenas uma variável por vez e apenas para algumas das classes – as predições feitas pela Regressão Linear não são visivelmente melhores do que a distribuição ao acaso.

16,67% corresponde aos níveis de acerto de um dado não viciado. Esta constatação corresponde à avaliação de sua performance, como inferido pela análise visual do seu *scatter plot* (Gráf. 13).

Gráfico 13 – *Scatter plot*, *LinReg*: *Effect*(*Mean*) *vs* Incid.; Hospit.; Letalid.



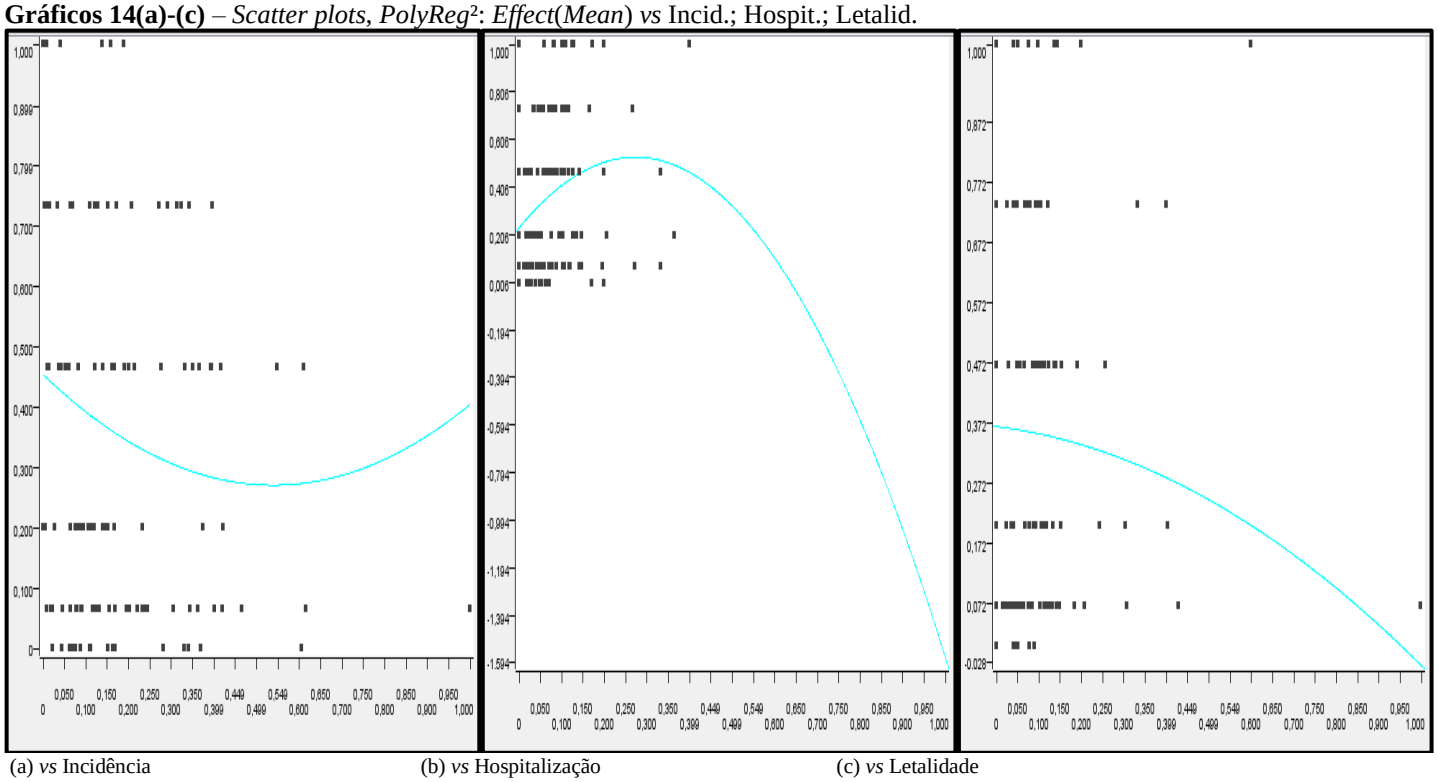
Em uma sequência lógica à R. Linear procedeu-se aqui à análise da Regressão Polinomial em grau 2 (*PolyReg*²), *i.e.*, a que busca a curva quadrática que melhor se ajuste aos pontos de dados. Considera-se a Linear como um caso particular das Polinomiais, para grau =1. De modo similar à *LinReg*, as análises foram feitas usando-se a MC (Tab. 16).

Tabela 16 – Matriz de Confusão da *PolyReg*²

Effect (Mean)Class(S) \ Prediction(Mean)Class(S)	0.8	0.6	0.4	0.2	0.1	0.05
0.8	0	7	1	0	0	0
0.6	1	5	4	0	0	0
0.4	0	1	6	0	0	0
0.2	0	3	8	0	0	0
0.1	0	2	9	2	0	0
0.05	0	0	5	0	0	0
Correct classified: 11			Wrong classified: 43			
Accuracy: 20,37%			Error: 79,63%			

A análise da MC acima e, abaixo, dos scatter plots da PolyReg² mostra que, para a maior parte das classes de resposta, das observações e das variáveis – exceto em algumas delas, e geralmente para apenas uma variável por vez e apenas para algumas das classes – as predições feitas pela PolyReg² não são perceptivelmente melhores do que a distribuição ao acaso, o que corresponde à avaliação da performance do algoritmo inferida a partir da análise visual dos Scatter plots mostrado nos Gráfs. 14(a)-(c).

Gráficos 14(a)-(c) – *Scatter plots, PolyReg*²: *Effect*(*Mean*) *vs* Incid.; Hospit.; Letalid.



(a) *vs* Incidência (b) *vs* Hospitalização (c) *vs* Letalidade

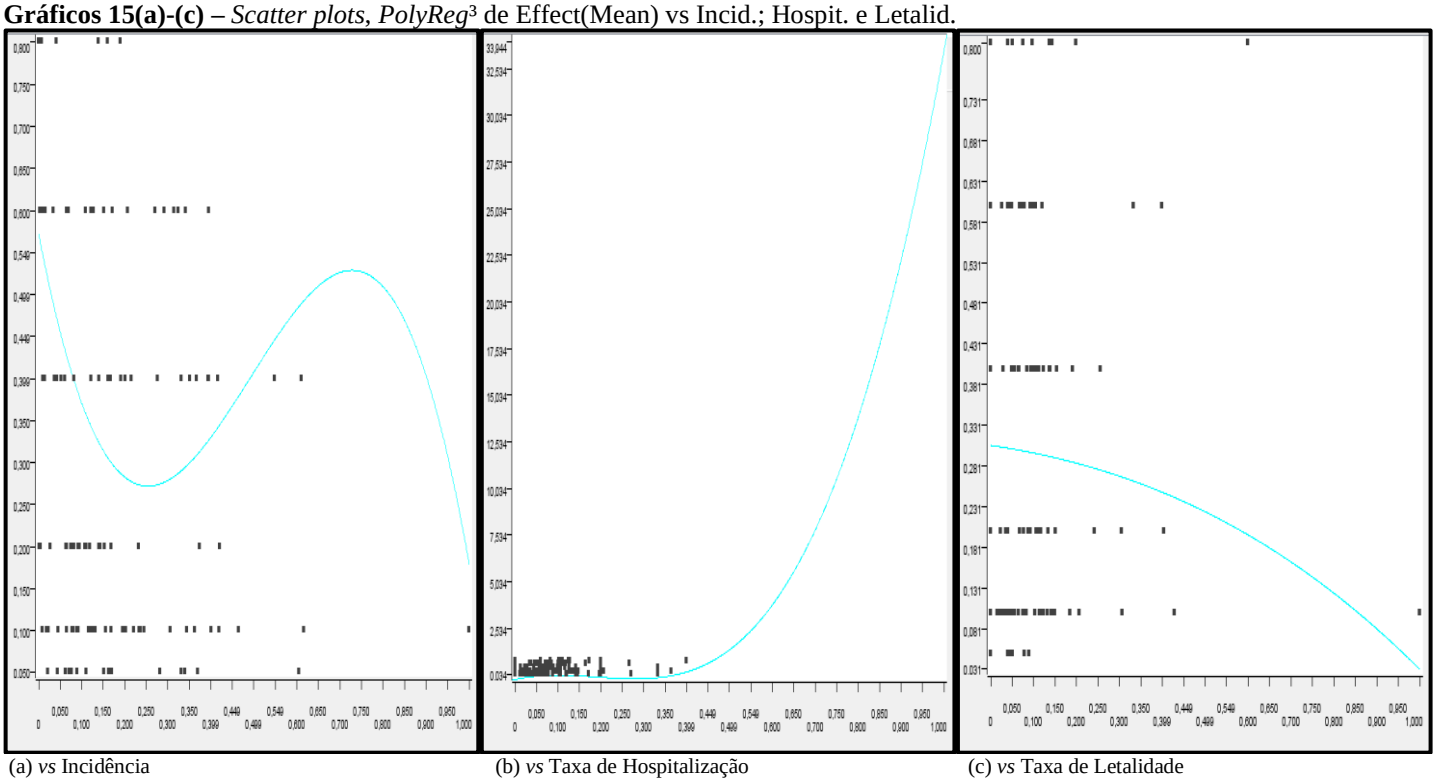
Em uma sequência lógica às regressões anteriores, procede-se aqui à análise da Regressão Polinomial em grau 3 (*PolyReg*³), *i.e.*, que busque a curva cúbica que melhor se ajuste aos dados. De modo similar à *PolyReg*², foram feitas as análises pela MC e pelas curvas de distribuição (respectivamente, na Tab. 17 e nos Grafs. 15(a)-(c)).

Tabela 17 – Matriz de Confusão da *PolyReg*³

Effect (Mean)Class(S) \ Prediction(Mean)Class(S)	0.8	0.6	0.4	0.2	0.1	0.05
0.8	3	3	1	0	0	0
0.6	3	4	2	0	0	0
0.4	0	1	5	0	0	0
0.2	0	4	5	2	0	0
0.1	1	2	9	1	0	0
0.05	0	2	3	0	0	0
Correct classified: 14			Wrong classified: 37			
Accuracy: 27,451%			Error: 72,549%			

A análise da MC acima e, abaixo, dos scatter plots, mostra que, pela análise da acurácia e da comparação entre as métricas reais e as preditas, e para a maior parte das observações e das variáveis – exceto, em algumas delas, geralmente para apenas uma variável por análise e para algumas das classes – as predições feitas pela PolyReg³ não são perceptivelmente melhores do que a distribuição ao acaso, com uma acurácia pouco superior a uma predição acertada a cada três equivocadas.

Gráficos 15(a)-(c) – *Scatter plots*, *PolyReg*³ de Effect(Mean) vs Incid.; Hospit. e Letalid.



(a) *vs* Incidência (b) *vs* Taxa de Hospitalização (c) *vs* Taxa de Letalidade

E a análise imediatamente subsequente é a da possível associação entre as variáveis usando o algoritmo da Regressão Polinomial de grau 4 (*PolyReg*^{4}), *i.e.*, a análise que busca a curva de quarta potência que melhor se ajuste aos pontos dos dados.

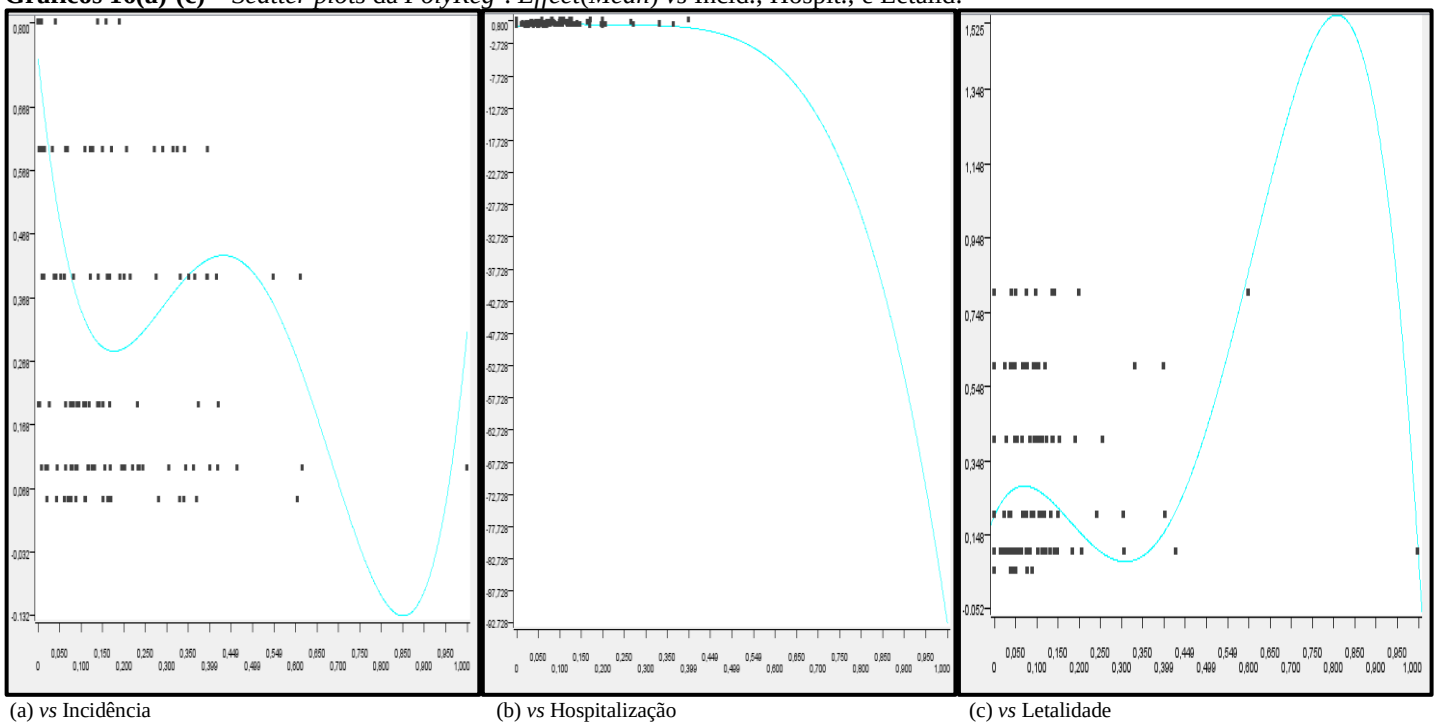
Após a configuração dos parâmetros e a execução do trecho do workflow correspondente à *PolyReg*^{4}, foi gerada a Matriz de confusão deste algoritmo (Tab. 18).

Tabela 18 – Matriz de Confusão da *PolyReg*[4]

Effect (Mean)Class(S) \ Prediction(Mean)Class(S)	0.8	0.6	0.4	0.2	0.1	0.05
0.8	4	2	1	0	0	0
0.6	4	1	3	1	0	0
0.4	0	0	5	1	0	0
0.2	0	5	4	2	0	0
0.1	1	1	10	1	0	0
0.05	0	4	1	0	0	0
Correct classified: 12			Wrong classified: 39			
Accuracy: 23,529%			Error: 76,471%			

A análise da acurácia na MC (e da comparação entre as métricas reais e as preditas nos scatter plots (Gráfs. 16(a)-(c)) mostra que, para a maior parte das observações e das variáveis – exceto, em algumas delas, e geralmente apenas para uma variável por vez, e apenas para algumas das classes – as predições feitas pela PolyReg4 não são perceptivelmente melhores do que a distribuição ao acaso, com uma acurácia pouco inferior a uma predição acertada a cada três equivocadas.

Gráficos 16(a)-(c) – *Scatter plots* da *PolyReg*[4]: *Effect*(*Mean*) *vs* Incid.; Hospit.; e Letalid.



(a) *vs* Incidência (b) *vs* Hospitalização (c) *vs* Letalidade

A aplicação de algoritmos de regressão prosseguiu para a Regressão Logística (*LogReg*). A *LogReg* é análoga às anteriores, e também *mutatis mutandis* diretamente comparável a elas, porém com algumas diferenças. A primeira, mais óbvia, é a de que trabalha não com potências, mas com logaritmos do seu argumento. Outra é a de que a variável dependente é categórica (ao contrário das Linear e Polinomial, que são numéricas). Embora

frequentemente seja aplicada à predição de probabilidades de uma variável categórica binária, também pode ser aplicada a variáveis multinomiais, caso este em que são calculadas as probabilidades para uma das classes (dita a de referência), para a comparação com as performances das demais classes. Outra diferença de sua aplicação em um workflow similar no Knime é uma consequência direta da anterior, a de que, ao invés de converter os tipos de dados de *double* para *string*, converte-se os valores das variáveis categóricas em numéricas (do tipo *double*), para os cálculos subsequentes. Ainda outra diferença deve-se à simplificação do workflow, pois não foram necessários os *nodes* para a conversão de valores médios calculados para as classes em que os dados são agrupados. A seguir, apresenta-se a Matriz de Confusão (MC) da *LogReg* na Tab. 19. Após, vêm o Apêndice D, que apresenta os gráficos das curvas ROC correspondentes, e a Tabela 22, que mostra o resumo das comparações das performances preditoras (por meio de suas respectivas AUCs).

Tabela 19 – Matriz de Confusão da *LogReg*

Effect (Mean)Class(S) \ Prediction((Effect (Mean))(S)	0.8	0.6	0.4	0.2	0.1	0.05
0.8	4	0	0	2	2	0
0.6	6	0	1	0	2	1
0.4	1	0	1	0	5	0
0.2	2	0	0	1	6	2
0.1	1	0	2	0	10	1
0.05	0	0	2	0	2	1
Correct classified: 17			Wrong classified: 38			
Accuracy: 30,909%			Error: 69,091%			

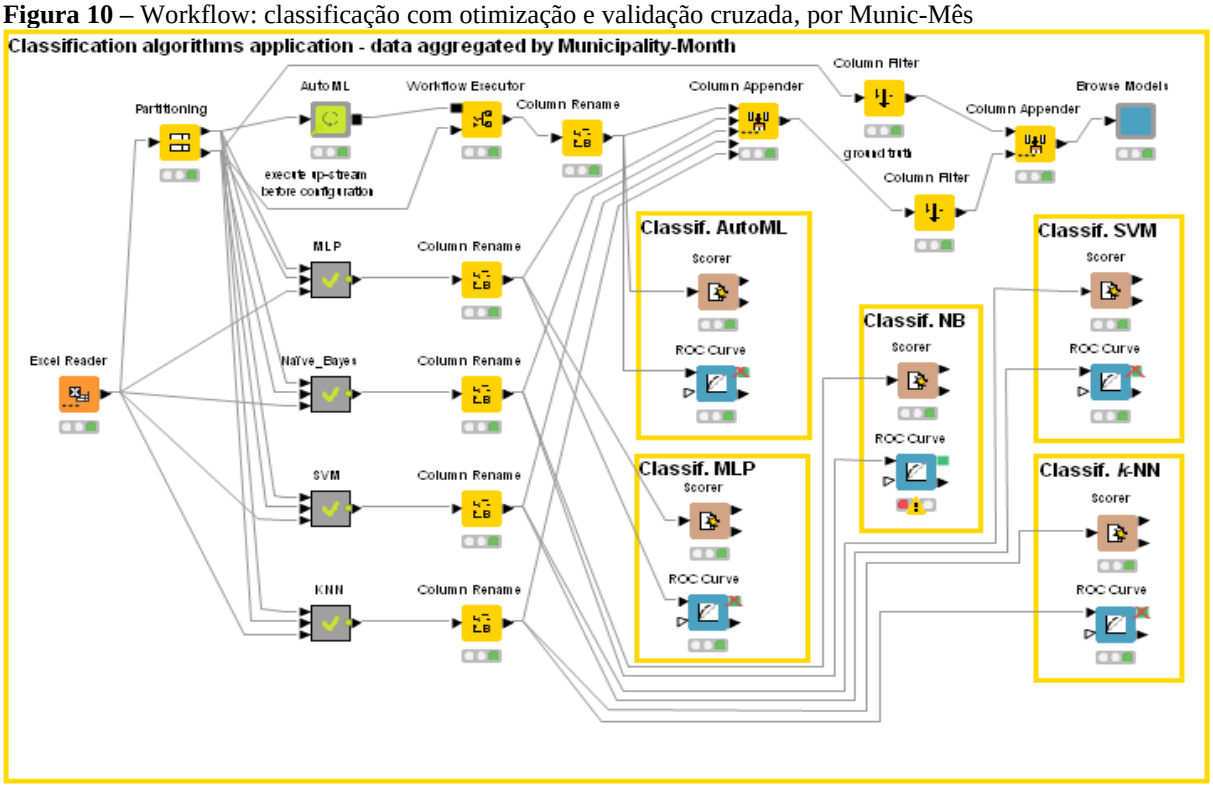
A análise da MC (acurácia de 30,91%) mostra que, pela comparação entre as métricas reais e as preditas, e para a maior parte das observações e das variáveis – exceto, em algumas delas, geralmente apenas para uma variável por vez e apenas para algumas das classes – as predições pela *LogReg* (para a métrica em função das três variáveis) não são perceptivelmente melhores do que a distribuição ao acaso, apresentando menos do que uma predição acertada a cada duas equivocadas.

Em resumo, para os dados trabalhados, nenhum modelo com qualquer dos cinco algoritmos de regressão, e para qualquer das suas classes, mostrou um padrão regular ou capacidade preditiva aceitável. Mas recorda-se, uma vez mais, que o objetivo é a demonstração da construção e da verificação da operacionalidade de um modelo e da forma de extração de conhecimento a partir dele sobre as associações entre os valores das diferentes variáveis e em diferentes circunstâncias, com diferentes algoritmos disponíveis no Knime.

4.5.3 Predições com a classificação: *k*-NN, SVM, Naïve Bayes, MLP e AutoML

Foi feita uma comparação entre a aplicação dos algoritmos *Support Vector Machines* (SVM), *k Nearest Neighbors* (*k*-NN), Naïve Bayes, *MultiLayer Perceptrons* (MLP) e para a classificação das respostas em um dos intervalos representados pelas seis classes de trabalho, conforme representado no trecho do workflow (Fig. 10). Este processo foi realizado com uma aplicação otimizada dos algoritmos recorrendo à validação cruzada. Para operacionalizar este procedimento sem, no entanto, deixar o trecho correspondente do workflow demasiado carregado de diferentes *nodes* e suas interconexões, foram usados *metanodes* e *components* (sendo que é nestes que foi feito o aninhamento de operações mais complexas e a ferramenta adicional da validação cruzada). Deste modo, foi possível testar incremental e iterativamente um conjunto de parâmetros (no Knime, tratados como *flow variables*) para cada *node* ou sequência de *nodes* aninhados, dentro de um intervalo de valores (definido pelo pesquisador) como compatíveis com a realidade estimada para a questão investigada.

Figura 10 – Workflow: classificação com otimização e validação cruzada, por Munic-Mês



Os resultados da aplicação destes algoritmos podem ser observados em maior LOD – ou "*drill down*" – ao final das seções subsequentes, que são dedicadas ao detalhamento de cada um dos algoritmos e à "abertura" (de alguns) dos *metanodes* e *components* contidos nele. Os

resultados das predições usando as regressões e classificações são apresentados na Tab. 22, que compara os parâmetros usados e as correspondentes performances obtidas.

Os modelos apresentados nesta dissertação tiveram seus resultados comparados entre si, com a finalidade de mostrar uma das possibilidades de procedimento de seleção entre dois ou mais algoritmos de classificação. Porém, recorda-se aqui que os procedimentos de execução e de comparação propriamente ditos (e não propriamente o conjunto dos valores obtidos) é que atendem aos objetivos da pesquisa, uma vez que a massa de dados foi demasiado diminuta, o que facilmente (se não previsivelmente) poderia estar distorcida ou enviesada, ou, ao contrário, distorcer ou enviesar os resultados da aplicação dos algoritmos. Os resultados e conhecimento deles extraídos podem facilmente ser devidos ao acaso ou a fatores espúrios, de modo que os valores específicos obtidos foram tomados como exemplo da construção e design do workflow e do encadeamento de operações. Porém estes valores carecem de representatividade e, assim, também as conclusões tiradas da comparação entre os resultados individuais (ou do conjunto) das rodadas[9] dos algoritmos.

4.5.4 Classificação dos dados com o algoritmo *k*-NN

São agora apresentados os resultados do treinamento do algoritmo *k*-NN para a predição do enquadramento de novos valores a partir de registros anteriores para os quais o algoritmo foi treinado.

Através da execução do *component* denominado *Browse Models* (localizado ao final do workflow), foi possível a seleção da (e o acesso à) Matriz de Confusão do *k*-NN (Fig. 11), a qual permite que seja avaliada a performance deste algoritmo, *i.e.*, o percentual das predições que corresponderam (ou melhor, coincidiram com) as métricas reais apuradas para as reduções nos fluxos dos consultórios:

[9] Adota-se o termo "rodada" como um ciclo completo de execução do algoritmo, até que ele encerre sua operação com base em critérios pré-estabelecidos durante seu design e configuração.

Figura 11 – Matriz de Confusão da aplicação do *k*-NN

Select Model

Prediction KNN

Scorer View

Confusion Matrix

	0.8 (Predict...	0.6 (Predict...	0.4 (Predict...	0.2 (Predict...	0.1 (Predict...	0.05 (Predic...	
0.8 (Actual)	1	1	0	1	3	0	16.67%
0.6 (Actual)	1	1	1	2	4	0	11.11%
0.4 (Actual)	0	0	2	4	3	0	22.22%
0.2 (Actual)	0	2	1	3	3	0	33.33%
0.1 (Actual)	1	2	0	4	8	0	53.33%
0.05 (Actual)	0	1	1	1	4	0	0.00%
	33.33%	14.29%	40.00%	20.00%	32.00%	undefined	

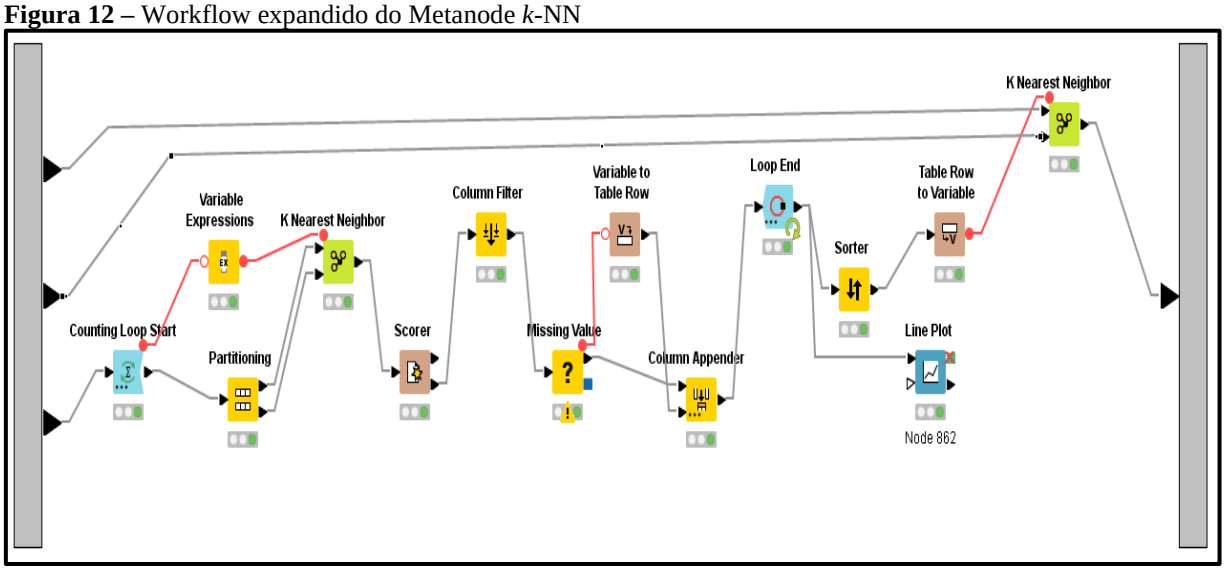
Overall Statistics

Overall Accuracy	Overall Error	Cohen's kappa (κ)	Correctly Classified	Incorrectly Classified
27.27%	72.73%	0.079	15	40

A acurácia geral foi 27,27% (ou de pouco mais de um quarto das predições acertadas). No entanto, a distribuição nas diferentes classes lembra um processo bastante errático.

O uso de um *metanode* no *k*-NN mostra (tal como em todos os algoritmos aqui trabalhados e apresentados) trechos de alguma extensão, o que, sem seu uso, tornaria menos imediata uma compreensão dos processos executados ao longo dos diferentes passos, etapas e fases do algoritmo (Fig. 12):

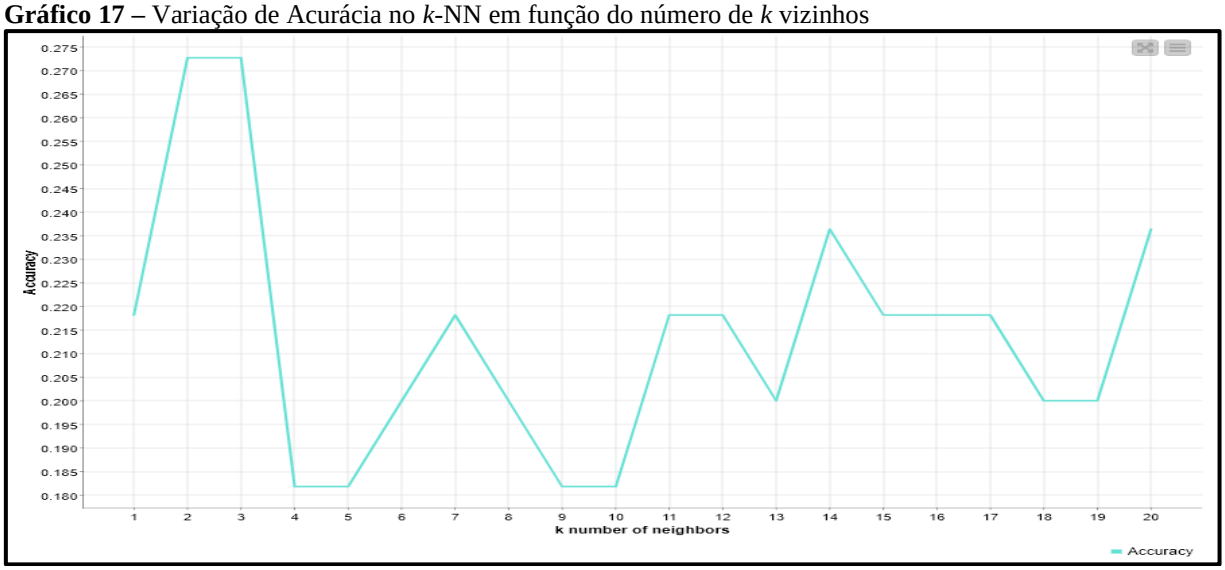
Figura 12 – Workflow expandido do Metanode *k*-NN



O *metanode* do *k*-NN, uma vez aberto, parece conter mais *nodes* do que os demais algoritmos, porém é bem menos complexo, por não apresentar *components*, e estes contêm outras cadeias de *nodes* (ou outros *components* aninhados).

Deste modo, para os sucessivos *k* vizinhos considerados (no intervalo [01; 20], conforme o Gráf. 17), as maiores acurácias, embora pequenas, atingem um máximo (= 0.273) para *k* = 2 e 3. Para as finalidades desta pesquisa, foi arbitrado que o número mais adequado é o de *k* = 3 (vizinhos mais próximos, NN).

Gráfico 17 – Variação de Acurácia no *k*-NN em função do número de *k* vizinhos

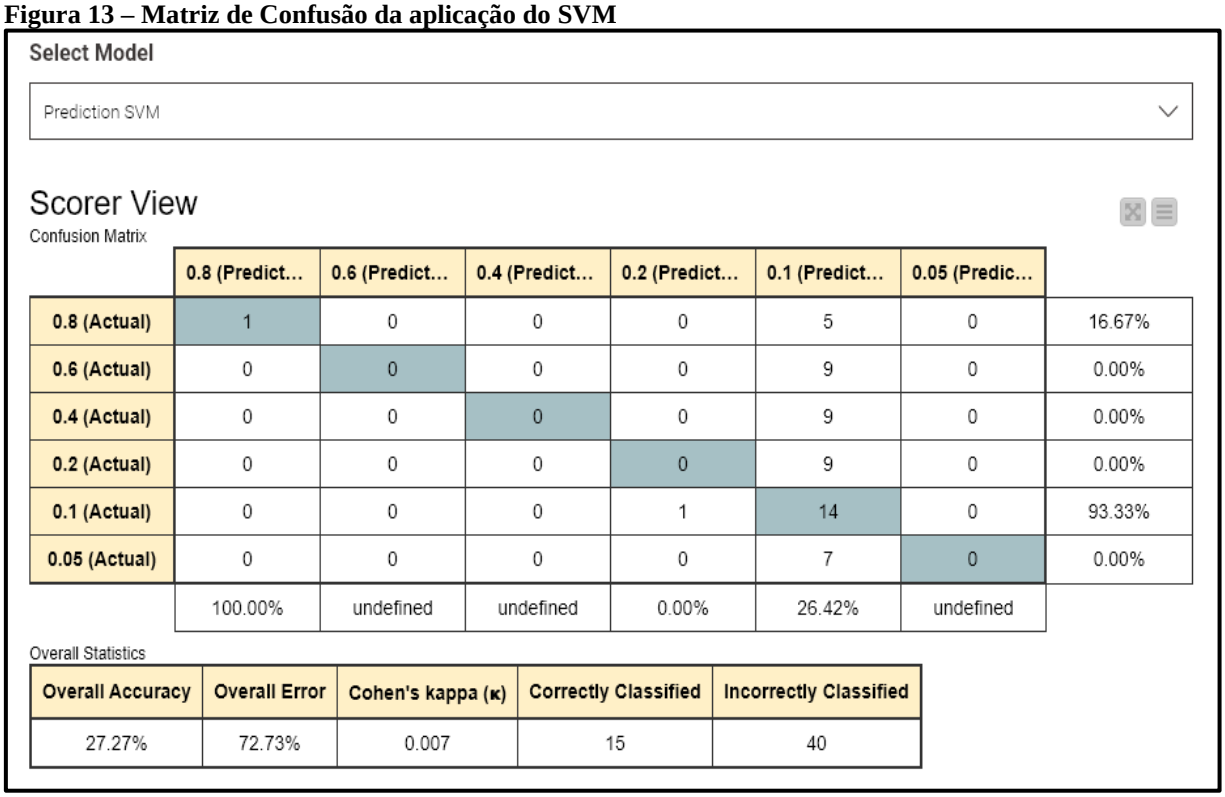


4.5.5 Classificação dos dados com o algoritmo SVM

São agora apresentados os resultados do treinamento do algoritmo das *Support Vector Machines* para a predição do enquadramento de novos valores a partir de registros anteriores para os quais o algoritmo tenha sido treinado.

Através da execução do *component* denominado *Browse Models* (localizado ao final do workflow), foi possível a visualização (Fig. 13) e análise, para a avaliação da performance do algoritmo SVM, *i.e.*, o percentual das predições que corresponderam (ou melhor, coincidiram com) as métricas reais apuradas para as reduções nos fluxos dos consultórios:

Figura 13 – Matriz de Confusão da aplicação do SVM



Select Model

Prediction SVM

Scorer View

Confusion Matrix

	0.8 (Predict...	0.6 (Predict...	0.4 (Predict...	0.2 (Predict...	0.1 (Predict...	0.05 (Predic...	
0.8 (Actual)	1	0	0	0	5	0	16.67%
0.6 (Actual)	0	0	0	0	9	0	0.00%
0.4 (Actual)	0	0	0	0	9	0	0.00%
0.2 (Actual)	0	0	0	0	9	0	0.00%
0.1 (Actual)	0	0	0	1	14	0	93.33%
0.05 (Actual)	0	0	0	0	7	0	0.00%
	100.00%	undefined	undefined	0.00%	26.42%	undefined	

Overall Statistics

Overall Accuracy	Overall Error	Cohen's kappa (κ)	Correctly Classified	Incorrectly Classified
27.27%	72.73%	0.007	15	40

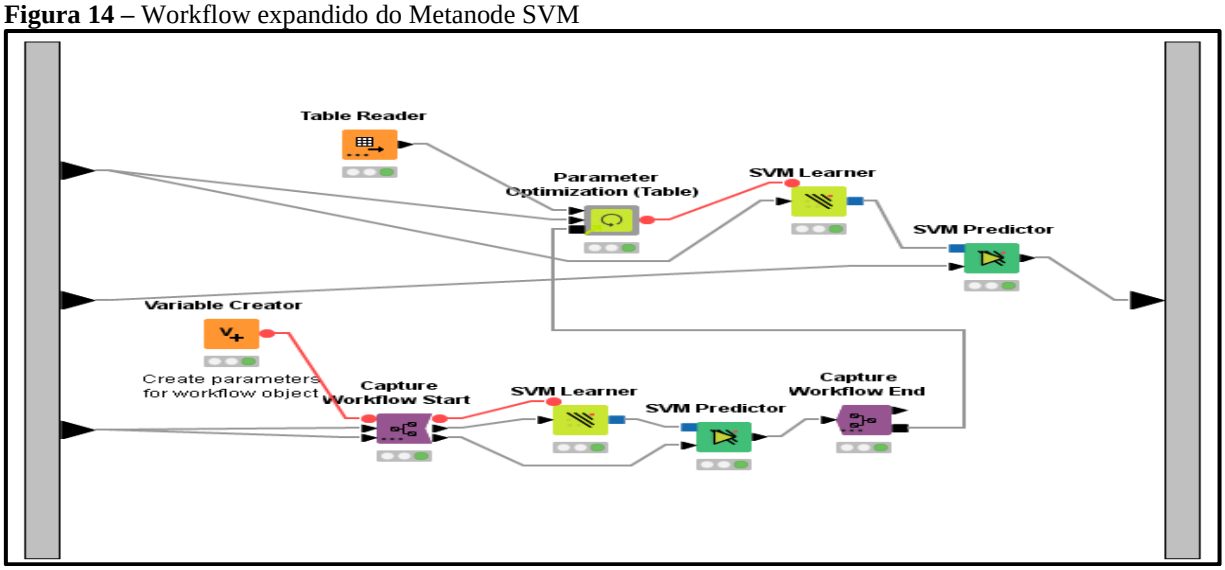
Este algoritmo (o SVM) teve sua performance avaliada (por sua acurácia geral) em 27,27% (ou pouco mais de uma quarta parte de predições corretas), predizendo acertadamente apenas 15 dentre 55 tentativas e errando as 40 tentativas remanescentes. E o modelo aqui construído permitiu esta avaliação com apenas um *click* para a execução do algoritmo SVM e a extração automatizada de sua performance, *i.e.*, o percentual de suas predições que corresponderam (ou melhor, coincidiram com) as métricas reais apuradas para as reduções nos fluxos dos consultórios. O desenvolvimento propriamente dito do modelo aqui apresentado foi uma tarefa (aparentemente) não muito simples, mas que pode ser replicada (com adaptações mínimas, se alguma) em inúmeras outras execuções das atividades de ML necessárias dentro de um ciclo de DS e para variados questões de pesquisa e *datasets*.

Destaca-se também a visível disparidade na distribuição das predições pelo algoritmo entre as diferentes classes, com praticamente a totalidade dos resultados da execução estando concentrados em apenas uma das classes, a “0.1”. E embora o valor da acurácia geral seja o mesmo do obtido com o *k*-NN (ver seção subsequente), a distribuição entre as diferentes classes é muito diferente entre ambos. Portanto, os resultados apresentados são, para (quase) todos os efeitos, marcadamente distintos entre ambos os algoritmos.

A “abertura” ou “expansão” (respectivamente, em uma guia separada ou na mesma área) de um destes WIMPs complexos, o do SVM, permitiu observar a adaptação feita os seguintes

trechos do workflow que, devido à opção pelo uso de *metanodes*, não estavam imediatamente visualizáveis (Fig. 14):

Figura 14 – Workflow expandido do Metanode SVM



No entanto, o *component* de otimização de parâmetros *Parameter Optimization (Table)*, por ser bem mais complexo e extenso, contendo os respectivos *metanodes* e *components* (nele aninhados e com os quais ele foi desenvolvido), uma vez expandido, gerou um trecho de workflow demasiado extenso para ser representado em uma imagem no corpo (ou mesmo em um dos Apêndices) do presente trabalho. Por esta perspectiva, podem ser tecidas algumas considerações sobre este *component*, como a de que, do ponto de vista dos usuários finais, ele pode ser descrito como um *Black-Box Classifier* – *i.e.*, do qual o usuário prescinde de uma compreensão muito aprofundada, ficando a cargo do desenvolvedor o design e a elaboração de procedimentos encadeados que tiverem maior envergadura e complexidade – ao qual é delegada boa parte da complexidade das etapas e cálculos sequenciais dentro do processo de classificação e comparação de performances. Esta limitação corresponde não somente à experiência comum e frequente dos pesquisadores, porém mesmo da vida urbana cotidiana, em que equipamentos, softwares e sistemas operacionais, composições químicas, rotinas de acesso e uso de serviços de diferentes tipos de empresas e instituições são apenas parcialmente compreendidas, porém regularmente usadas por boa parte de seus usuários via ferramentas automatizadas.

Entretanto, aos que desejarem obter uma compreensão mais aprofundada dos diferentes processos que estão ocorrendo sob a execução deste WIMP, a equipe do Knime elaborou uma detalhada descrição de toda a extensão *Knime Integrated Deployment*, em uma coleção de posts

minuciosa explicando estes processos e explorando algumas de suas possibilidades no já referido Blog do Knime.

4.5.6 Classificação dos dados com o algoritmo Naïve Bayes

São agora apresentados os resultados do treinamento do algoritmo Naïve Bayes (NB) para a predição do enquadramento de novos valores a partir de registros anteriores para os quais o algoritmo tenha sido treinado.

Através da execução do *component* denominado *Browse Models* (localizado ao final do workflow), foi possível a seleção da (e o acesso à) Matriz de Confusão do NB (Fig. 15), a qual permite que seja avaliada a performance:

Figura 15 – Matriz de Confusão da aplicação do Naïve Bayes

Select Model

Prediction Naive Bayes

Scorer View

Confusion Matrix

	0.8 (Predict…	0.6 (Predict…	0.4 (Predict…	0.2 (Predict…	0.1 (Predict…	0.05 (Predic…	
0.8 (Actual)	6	0	0	0	0	0	100.00%
0.6 (Actual)	0	9	0	0	0	0	100.00%
0.4 (Actual)	0	0	9	0	0	0	100.00%
0.2 (Actual)	0	0	0	8	1	0	88.89%
0.1 (Actual)	0	0	0	4	3	8	20.00%
0.05 (Actual)	0	0	0	0	0	7	100.00%
	100.00%	100.00%	100.00%	66.67%	75.00%	46.67%	

Overall Statistics

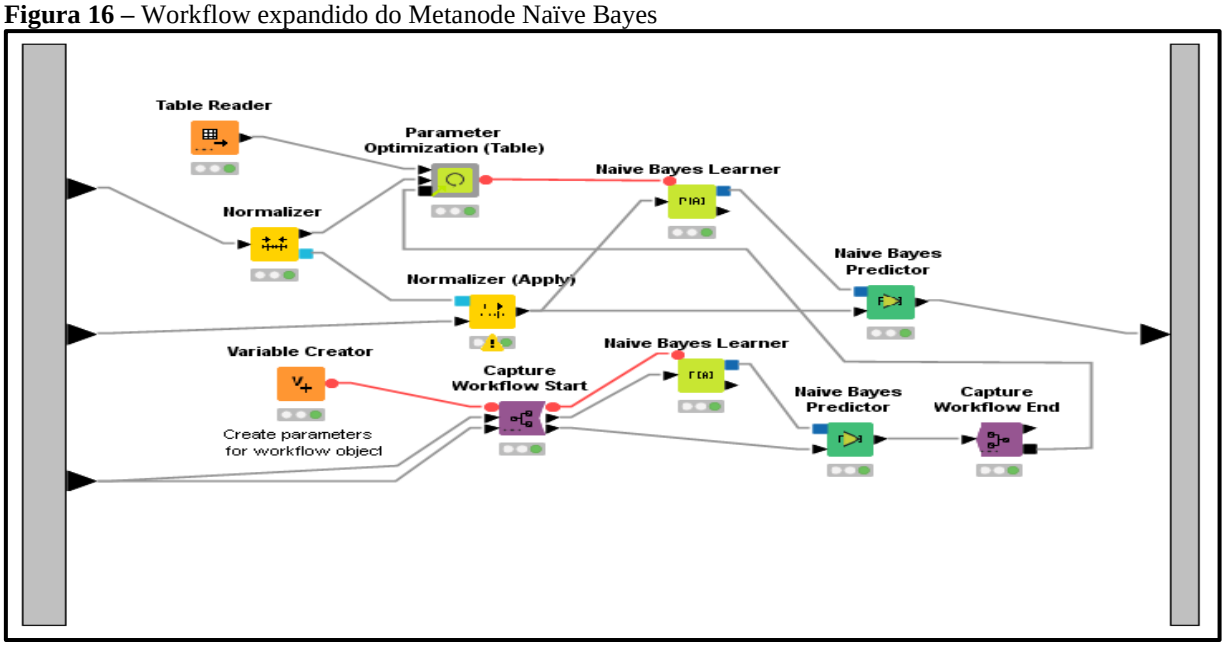
Overall Accuracy	Overall Error	Cohen's kappa (κ)	Correctly Classified	Incorrectly Classified
76.36%	23.64%	0.720	42	13

Este algoritmo (o NB) teve sua performance avaliada (por sua acurácia geral) em 76,36% (ou pouco mais de três quartos de predições corretas), predizendo acertadamente 42 dentre 55 tentativas e errando as 13 remanescentes. Este desempenho, i.e., o percentual das predições que corresponderam às (ou melhor, coincidiram com as) métricas reais apuradas para as reduções nos fluxos dos consultórios esteve marcado contraste com os resultados precedentes (os do k-NN e do SVM). Um dos resultados imediatamente aparente é a grande concentração das predições ao redor do eixo principal (diagonal) da MC. Ou com classificações em classes imediatamente adjacentes a estas. E o modelo aqui construído também permitiu esta avaliação

com apenas um click após a execução do algoritmo. Embora o desenvolvimento do modelo aqui apresentado tenha sido uma tarefa (aparentemente) não muito simples, ela pode ser replicada (com adaptações mínimas, e com mínimo esforço intelectual e manual) para inúmeras outras execuções das atividades de ML e diferentes conjuntos de dados.

A "abertura" ou "expansão" (respectivamente, em uma guia separada ou na mesma área) de um destes WIMPs complexos, o do NB (Fig. 16), permitiu observar a adaptação feita os seguintes trechos do workflow que não estavam imediatamente visualizáveis:

Figura 16 – Workflow expandido do Metanode Naïve Bayes



Tal como com os demais algoritmos otimizados por este workflow, a "abertura" ou "expansão" do *component* denominado *Parameter Optimization (Table)* é o mais extenso e foge às capacidades de representação no corpo (ou Apêndices) deste relatório de pesquisa. Porém mantém-se igualmente válida a recomendação de que o leitor busque maior aprofundamento e compreensão nas citações recomendadas durante a apresentação dos resultados do SVM.

4.5.7 Classificação dos dados com os Perceptrons Multicamadas (MLP)

O algoritmo de NNs Multi-Layer Perceptron (MLP) foi aplicado aos dados, tal como nos algoritmos anteriores.

Através da execução do *component* denominado *Browse Models* (localizado ao final do workflow), foi possível a seleção da (e o acesso à) Matriz de Confusão dos MLP (Fig. 17), a qual permite que seja avaliada a performance deste algoritmo, bem como seus parâmetros otimizados:

Figura 17 – Matriz de Confusão do MLP

Select Model

Prediction MLP

Scorer View

Confusion Matrix

	0.8 (Predict...	0.6 (Predict...	0.4 (Predict...	0.2 (Predict...	0.1 (Predict...	0.05 (Predic...	
0.8 (Actual)	6	0	0	0	0	0	100.00%
0.6 (Actual)	0	9	0	0	0	0	100.00%
0.4 (Actual)	0	0	9	0	0	0	100.00%
0.2 (Actual)	0	0	0	9	0	0	100.00%
0.1 (Actual)	0	0	0	1	14	0	93.33%
0.05 (Actual)	0	0	0	0	0	7	100.00%
	100.00%	100.00%	100.00%	90.00%	100.00%	100.00%	

Overall Statistics

Overall Accuracy	Overall Error	Cohen's kappa (κ)	Correctly Classified	Incorrectly Classified
98.18%	1.82%	0.978	54	1

A acurácia geral calculada foi de 98,18%, representando um exemplo de como se pode atingir predições muito boas a excelentes com a aplicação de um algoritmo, sendo que esta execução predisse acertadamente 54 classificações dentre as 55 executadas.

A execução destes MLP permitiu a descoberta dos melhores valores para *n* (n° de neurônios por camada) e para *l* (número de camadas ocultas) como *n* = 18 e *l* = 8, como os que produziram a melhor performance para o MLP, que foi de 0.933 (Tab. 20

Tabela 20 – Melhores parâmetros para a NN pelo MLP

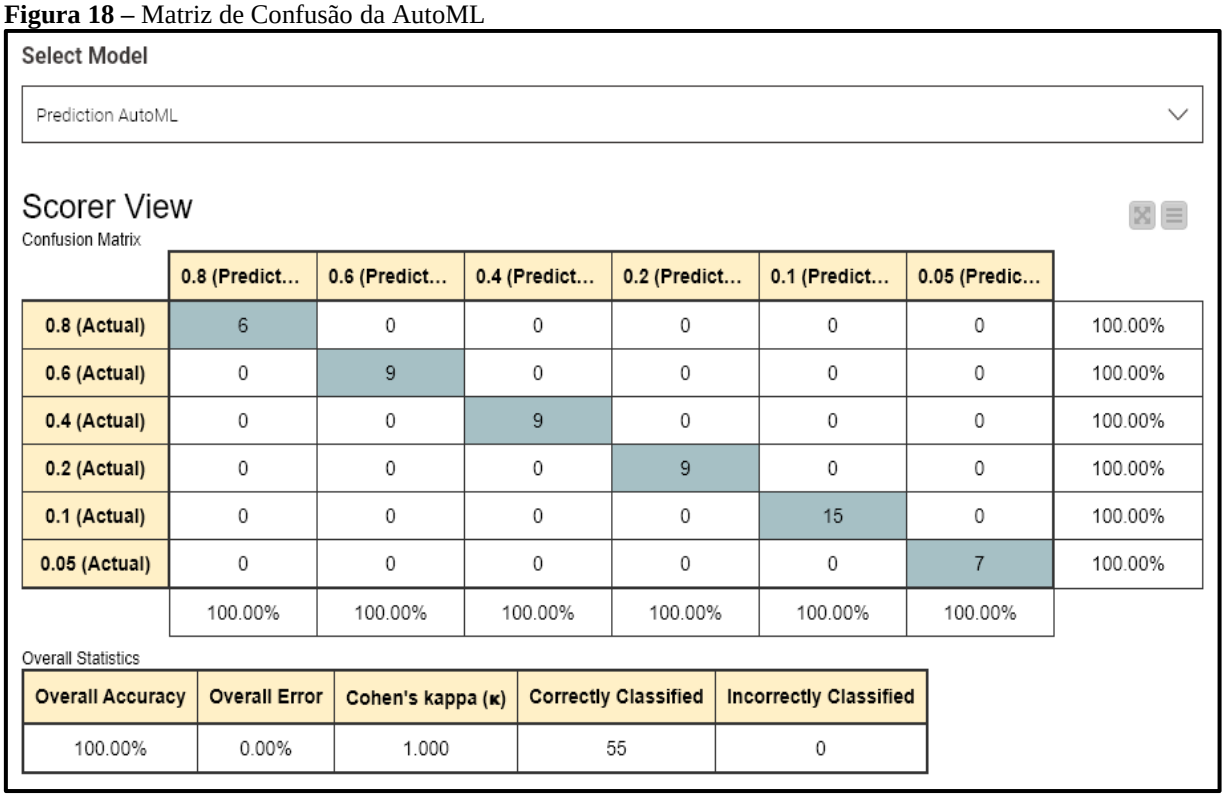
Row ID	MLP_max_number_iterations	MLP_number_hidden_layers	MLP_number_neurons_per_layer	Objective value
Best parameters	70	8	18	0.933

4.5.8 Classificação dos dados com a ML Automatizada (AutoML)

Esta seção mostra os resultados da aplicação dos dados à ML Automatizada (AutoML).

Através da execução do *component* denominado *Browse Models* (localizado ao final do workflow), foi possível a seleção da (e o acesso à) Matriz de Confusão da AutoML (Fig. 18), a qual permite que seja avaliada a performance deste algoritmo, *i.e.*, o percentual das predições que corresponderam (ou melhor, coincidiram com) as métricas reais apuradas para as reduções nos fluxos dos consultórios, bem como seus parâmetros otimizados:

Figura 18 – Matriz de Confusão da AutoML



	0.8 (Predict...	0.6 (Predict...	0.4 (Predict...	0.2 (Predict...	0.1 (Predict...	0.05 (Predic...	
0.8 (Actual)	6	0	0	0	0	0	100.00%
0.6 (Actual)	0	9	0	0	0	0	100.00%
0.4 (Actual)	0	0	9	0	0	0	100.00%
0.2 (Actual)	0	0	0	9	0	0	100.00%
0.1 (Actual)	0	0	0	0	15	0	100.00%
0.05 (Actual)	0	0	0	0	0	7	100.00%
	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	

Overall Accuracy	Overall Error	Cohen's kappa (κ)	Correctly Classified	Incorrectly Classified
100.00%	0.00%	1.000	55	0

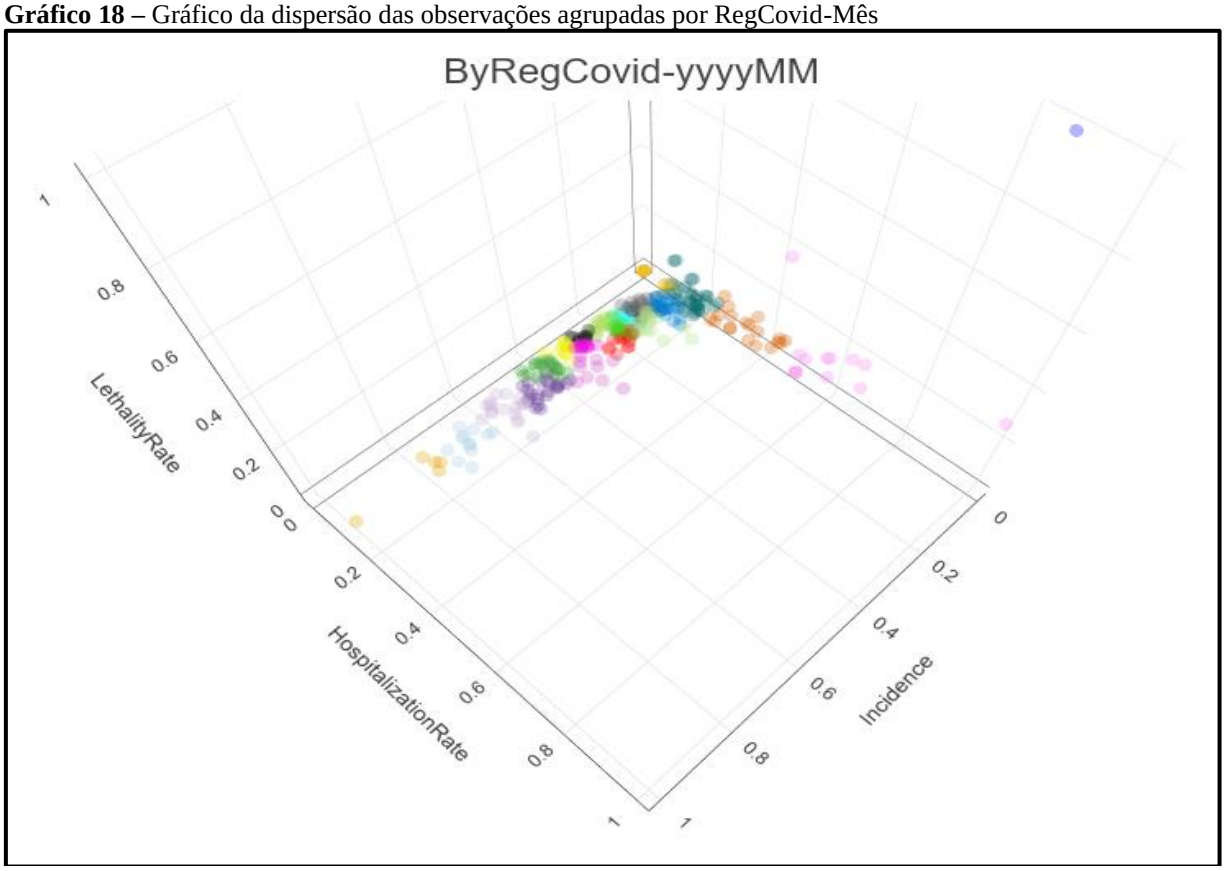
A acurácia geral calculada foi de 100%, representando um exemplo de como se pode atingir predições excelentes, e especialmente com um mínimo esforço para a configuração dos diferentes parâmetros e demais elementos que integram este algoritmo. O algoritmo predisse acertadamente todas as 55 classificações. Esta performance (ou desempenho) corresponde ao percentual das predições que corresponderam ao (ou melhor, coincidiram com) as métricas reais apuradas para as reduções nos fluxos dos consultórios, *i.e.*, representando a totalidade das predições.

4.5.9 Comparação entre *clusters* por Munic-Mês e por RegCovid-Mês

Conforme anunciado na Seção 4.3, foi feita uma comparação entre ambos os níveis de granularidade para a clusterização das observações.

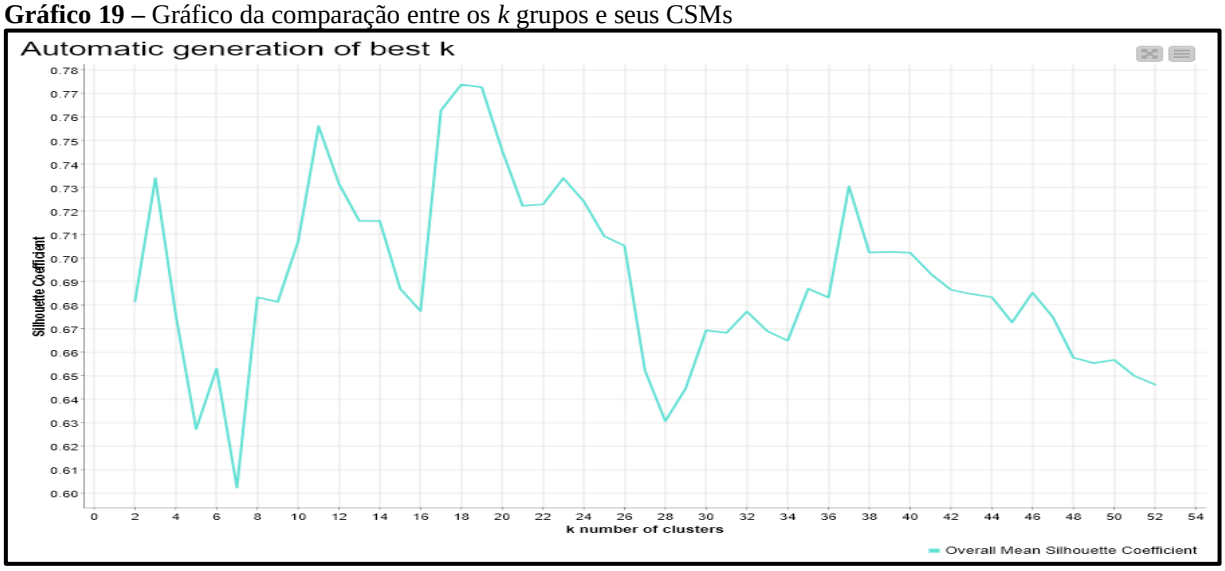
Os pontos, agrupados por RegCovid-Mês, correspondeu aos dados coletados, porém agrupados nas 21 Regiões Covid ao longo dos 16 meses de pesquisa, o que resultou em 333 observações – excetuando-se as 3 regiões que não apresentaram registros de casos em algum mês, o que ocorreu exclusivamente nos meses de março e abril de 2020, os mais próximos do início da pandemia. Aplicando-se a clusterização exclusivamente a partir de critérios de proximidade geográfica, e no caso desta pesquisa, usando o mesmo critério de agrupamento em 21 Regiões Covid, foi gerado um *scatter plot* mostrando, de modo similar ao já feito no agrupamento por Munic-Mês, com os 21 *clusters* mostrados (por cor) no Gráf. 18.

Gráfico 18 – Gráfico da dispersão das observações agrupadas por RegCovid-Mês



Este nível de granularidade apresenta a distinção visual entre os grupos de observações pelas dimensões das taxas de Hospitalização e de Letalidade, pois os pontos, exceto para baixas incidências, assumem valores baixos. Então, procedeu-se à otimização deste agrupamento. Segundo esta, que é apresentada na Fig.8, o trecho correspondente à otimização gerou um número $k = 3$ grupos de observações por RegCovid, mostrado no Gráf. 19.

Gráfico 19 – Gráfico da comparação entre os *k* grupos e seus CSMs



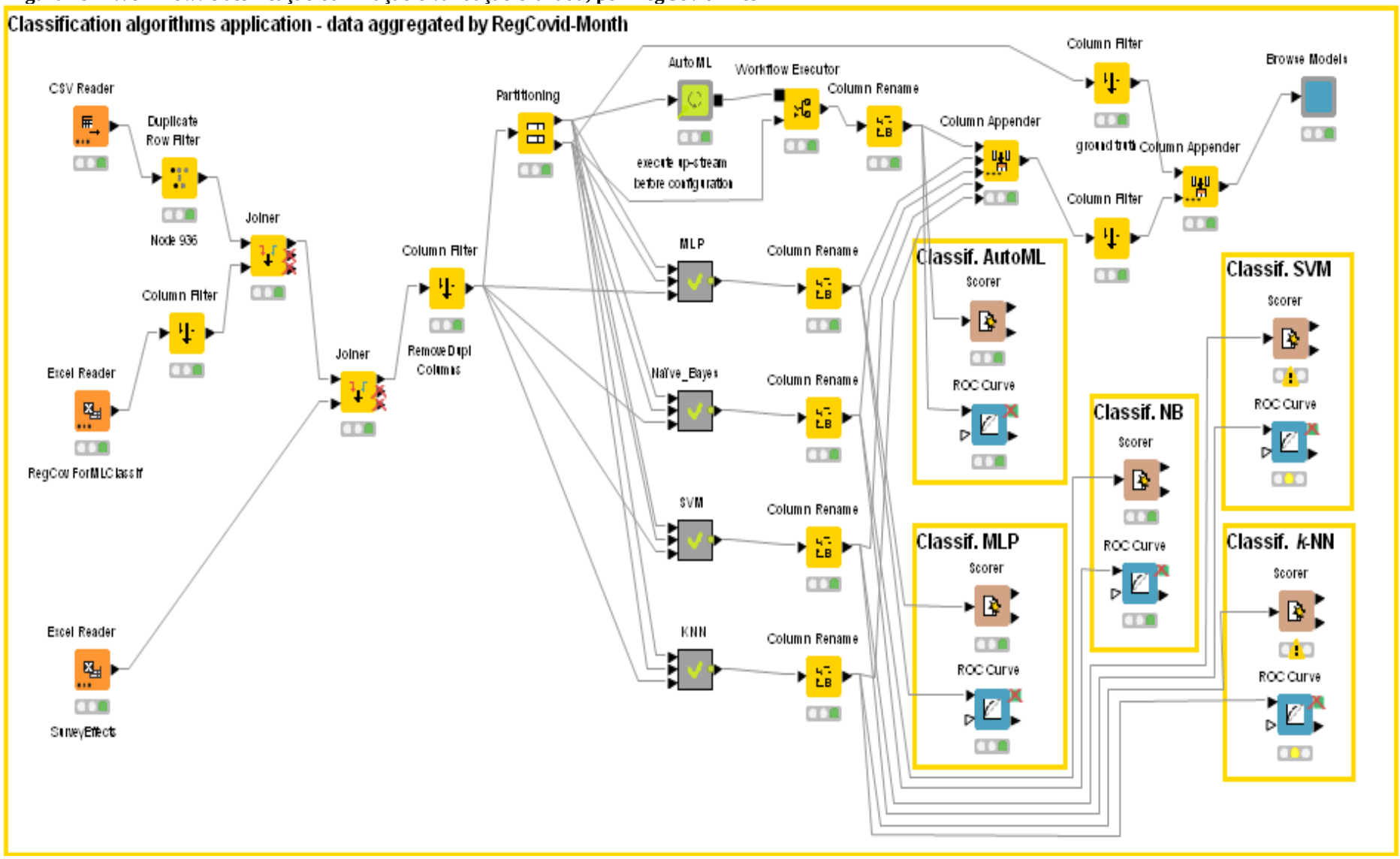
Pelo *Elbow Method*, o melhor número para a clusterização de observações agregadas conforme RegCovid-Mês é o de *k* = 3, pois corresponde à primeira inflexão no gráfico.

Este novo número otimizado de *clusters* é muito similar ao dos agrupamentos feitos com base em Munic-Mês (que foi de *k* = 4). No entanto, uma análise da comparação entre ambos os gráficos indica que as similaridades intragrupo e dissimilaridades intergrupos, *i.e.*, os respectivos graus de coesão dos e distanciamento entre os grupos (medido pelos seus CSMs) foram menores quando as observações foram agrupadas por município e mês do que quando por Região Covid e mês (Tabs. 19(a)-(b)).

A etapa subsequente foi a adaptação, execução dos mesmos algoritmos de classificação e a comparação de seus resultados com os dos que rodaram com agrupamentos por Munic-Mês.

A adaptação dos algoritmos de classificação para o agrupamento por RegCovid-mês foi mínima, alterando-se as DBs consultadas para a execução das *queries* dos algoritmos, o que é mostrado na Fig. 19).

Figura 19 – Workflow: classificação otimização e validação cruzada, por RegCovid-Mês



Os resultados da comparação entre os resultados da ML por Munic-Mês e por RegCovid-Mês são mostrados na Tab. 21.

Tabela 21 – Comparação dos resultados da ML por Munic-Mês e por RegCovid-Mês

Acurácia dos algoritmos	**Agrupamento por Munic-Mês**	**Concentra predições**	**Agrupamento por RegCovid-Mês**	**Concentra predições**
Melhor *k*-Means	4	-	3	-
SVM	27,27%	Na Classe 0,8	56,90%	Na Classe 0,8
k-NN	27,27%	Dispersão aleatória	56,90%	Na Classe 0,8
Naïve Bayes	76,36%	Na diagonal principal	98,28%	Na diagonal principal
MLP	98,18%	Na diagonal principal	100%	Na diagonal principal
AutoML	100%	Na diagonal principal	100%	Na diagonal principal

E a Tabela 22 apresenta um levantamento feito com a análise das Curvas ROC (e respectivas AUCs) comparando as:

a) métricas reais (levantadas durante a *survey*) com as três taxas selecionadas da COVID-19;
b) predições feitas pelos algoritmos com as três taxas selecionadas da COVID-19; e
c) métricas reais (levantadas durante a *survey*) das variações nos fluxos dos consultórios com as predições feitas pelos algoritmos para estas variações.

Estas comparações aplicam a análise visual das curvas ROC para identificar se há (ou não) associações aparentes o suficiente entre as variáveis (supostamente) preditoras e os efeitos observados (de retração no mercado-alvo da pesquisa).

Tabela 22 – Comparação entre performances dos algoritmos pelas AUCs

ALGORITHM	**Rates****	**Metrics *vs* Rates**						**Predictions *vs* Rates**						**Predictions *vs* Metrics**					
Rates* *vs* Classes		**0.8**	**0.6**	**0.4**	**0.2**	**0.1**	**0.05**	**0.8**	**0.6**	**0.4**	**0.2**	**0.1**	**0.05**	**0.8**	**0.6**	**0.4**	**0.2**	**0.1**	**0.05**
LinReg	Incid.	0.168	0.280	0.602	0.489	**0.714**	**0.812**	0.036	0.146	**0.927**	n/pred	n/pred	n/pred	**0.778**	**0.691**	0.259	n/pred	n/pred	n/pred
	Hospit.	**0.742**	0.628	0.533	0.411	0.278	0.516	**1.000**	**0.913**	0.017	n/pred	n/pred	n/pred						
	Lethalid.	0.537	0.427	0.617	0.514	0.459	0.482	0.379	0.253	**0.721**	n/pred	n/pred	n/pred						
PolyReg^2	Incid.	0.168	0.28	0.602	0.489	**0.714**	**0.812**	n/pred	0.108	**0.771**	**0.894**	n/pred	n/pred	n/pred	**0.672**	0.405	0.202	n/pred	n/pred
	Hospit.	**0.742**	**0,628**	0.533	0.411	0.278	0.516	n/pred	**0.947**	0.188	0.067	n/pred	n/pred						
	Lethalid.	0.537	0.427	0.617	0.514	0.459	0.482	n/pred	0.4	0.638	0.25	n/pred	n/pred						
PolyReg^3	Incid.	0.193	0.263	0.593	0.477	**0.707**	**0.808**	0.044	0.310	**0.708**	0.603	n/pred	n/pred	**0.701**	**0.777**	0.266	0.360	n/pred	n/pred
	Hospit.	**0.705**	0.643	0.545	0.421	0.285	0.527	**1.000**	**0.723**	0.337	0.047	n/pred	n/pred						
	Lethalid.	0.595	0.420	0.610	0.505	0.447	0.473	0.515	0.463	0.574	0.303	n/pred	n/pred						
PolyReg^4	Incid.	0.169	0.265	**0.681**	0.451	**0.691**	**0.800**	0.080	0.114	**0.832**	0.62	n/pred	n/pred	**0.883**	**0.759**	0.210	**0.511**	n/pred	n/pred
	Hospit.	**0.737**	0.620	0.48	0.442	0.3	0.550	0.574	**0.885**	0.343	0.213	n/pred	n/pred						
	Lethalid.	0.596	0.354	**0.704**	0.507	0.446	0.474	0.314	0.495	0.529	0.587	n/pred	n/pred						
LogReg	Incid.	0.178	0.296	0.607	0.501	**0.670**	**0.816**	0.054	n/pred	**0.779**	0.385	**0.816**	0.312	**0.788**	n/pred	0.294	**0.782**	0.345	0.372
	Hospit.	**0.726**	0.614	0.522	0.402	0.329	0.506	**0.855**	n/pred	**0.760**	**0.705**	0.228	0.076						
	Lethalid.	0.545	0.434	0.624	0.522	0.436	0.490	0.312	n/pred	0.536	**0.699**	**0.724**	0.090						
SVM	Incid.	0.344	0.413	0.435	0.431	**0,670**	**0.816**	0.000	n/pred	n/pred	**0.759**	**0.781**	**0.781**	**0.954**	n/pred	n/pred	0.432	0.392	0.080
	Hospit.	**0.697**	**0.563**	**0.652**	0.345	0.329	0.506	**0.722**	n/pred	n/pred	0.333	0.307	0.307						
	Lethalid.	0.347	**0.511**	**0.693**	0.533	0.436	0.490	0.102	n/pred	n/pred	0.389	0.517	0.517						
***k*-NN**	Incid.	0.344	0.413	0.435	0.431	**0.553**	**0.815**	0.109	0.207	**0.512**	0.457	**0.743**	n/pred	**0.683**	0.469	**0.532**	0.259	0.450	n/pred
	Hospit.	**0.697**	**0.563**	**0.652**	0.345	0.455	0.333	**0.750**	**0.688**	**0.612**	**0.573**	0.268	n/pred						
	Lethalid.	0.347	0.511	**0.693**	**0.533**	**0.554**	0.246	0.263	**0.649**	0.458	**0.571**	0.440	n/pred						
Naïve Bayes	Incid.	0.344	0.413	0.435	0.431	**0.553**	**0.815**	0.344	0.413	0.435	0.424	0.113	**0.878**	**1.000**	**0.870**	**0.674**	0.384	0.304	0.047
	Hospit.	**0.697**	**0.563**	**0.652**	0.345	0.455	0.333	**0.697**	**0.563**	**0.652**	0.397	**0.735**	0.263						
	Lethalid.	0.347	**0.511**	**0.693**	**0.533**	**0.554**	0.246	0.347	**0.511**	**0.693**	**0.527**	**0.777**	0.317						
MLP(RPROP)	Incid.	0.344	0.413	0.435	0.431	**0.553**	**0.815**	0.344	0.413	0.435	0.383	**0.597**	**0.815**	**1.000**	**0.870**	**0.674**	0.451	0.183	0.000
	Hospit.	**0.697**	**0.563**	**0.652**	0.345	0.455	0.333	**0.697**	**0.563**	**0.652**	0.418	0.406	0.333						
	Lethalid.	0.347	**0.511**	**0.693**	**0.533**	**0.554**	0.246	0.347	**0.511**	**0.693**	0.482	**0.594**	0.246						
AutoML	Incid.	0.344	0.413	0.435	0.431	**0.553**	**0.815**	0.344	0.413	0.435	0.431	**0.553**	**0.815**	**1.000**	**0.870**	**0.674**	0.478	0.175	0.000
	Hospit.	**0.697**	**0.563**	**0.652**	0.345	0.455	0.333	**0.697**	**0.563**	**0.652**	0.345	0.455	0.333						
	Lethalid.	0.347	**0.511**	**0.693**	**0.533**	**0.554**	0.246	0.347	**0.511**	**0.693**	**0.533**	**0.554**	0.246						

Notas: * Estão destacadas **em negrito** as AUCs de performance melhor do que a distribuição ao acaso, *i.e.*, > 0,5
** Em "Rates" (ou Taxas): "Incid" = Incidência; "Hospit" = Taxa de Hospitalização; e "Lethalid" = Taxa de Letalidade
n/pred = Sem predições pelos algoritmos

5 DISCUSSÃO DOS RESULTADOS

Este capítulo apresenta a discussão feita a partir dos resultados mostrados no capítulo precedente.

5.1 CARACTERÍSTICAS EPIDEMIOLÓGICAS DA COVID-19 NO RS

A COVID-19 apresentou as seguintes distribuições ao longo de seus primeiros 16 meses de vigência no RS:

a) Os números relativos aos casos registrados no RS apresentaram flutuações quase síncronas e de magnitude comparável à das taxas médias mundiais, replicando o "padrão de ondas" observado em outros locais ou países, *i.e.*, a doença reproduziu localmente um padrão de disseminação muito similar ao médio mundial;
b) Os diagnósticos positivos para COVID-19 foram atribuídos (principalmente) com base em Testes Rápidos (47%) ou de RT-PCR (50%), o que sugere uma distribuição muito similar entre ambas as modalidades de testagem; e esta constatação, isoladamente, pode representar a conjuntura nacional da época para a aquisição e distribuição de *kits* de testagem;
c) A disseminação da COVID-19, *i.e.*, a incidência de novos casos ao longo de todo o período da pesquisa, teve uma margem regular e consistentemente maior (embora pequena), no gênero feminino, para as diferentes faixas etárias e para os diferentes meses da pesquisa;
d) porém o gênero masculino apresentou, também ao longo de todo o período da pesquisa, pequenas, regulares e consistentemente maiores parcelas de pacientes com sinais e sintomas graves, *i.e.*, aqueles que levavam à internação hospitalar, para as diferentes faixas etárias e para os diferentes meses da pesquisa;
e) o gênero masculino também mostrou, ao longo de todo o período da pesquisa, pequenas, regulares e consistentemente maiores parcelas de pacientes com êxito fatal, *i.e.*, maiores parcelas de óbitos, em relação ao feminino;
d) A infecção pela doença mostrou uma marcada associação entre casos novos e os contingentes populacionais locais, caracterizando uma transmissão sustentada inter-humanos;
e) Os valores da variável SRAG (o registro de casos sintomáticos graves) apresenta uma associação estatística perfeita com os valores dos registros de internações, o que indica que os casos foram registrados como "SRAG" (quase exclusivamente) para os pacientes que foram hospitalizados. Portanto, parece ser suficientemente acertada a equivalência que foi assumida entre as variáveis "SRAG" e "Hospitalização";
f) A faixa etária dos 30 a 39 anos foi a que apresentou maiores contagens de novos casos, porém as faixas de 50 a 69 anos mostraram as maiores contagens de hospitalizações e as de acima de 60 anos mostraram as maiores mortalidades;
g) Foram analisados os índices de correlação entre as três taxas (Incidência, de Hospitalização e de Letalidade), para avaliar se eram interdependentes – circunstância que, caso fosse constatada, invalidaria a aplicação de alguns dos algoritmos que tiverem o pressuposto da independência das variáveis) – e estas taxas se mostraram não correlacionadas.

5.2 CLUSTERIZAÇÃO POR MUNICÍPIO-MÊS (*via* CSMs)

O agrupamento dos dados pelo k-Means – otimizado pelo Elbow Method com o uso dos CSMs – por similaridade e segundo as dimensões de análise, permitiu a identificação de quatro agrupamentos tão coesos e distintos entre si quanto possível, e mais do que com outras opções para número (naturais) de subgrupos.

Este processo de *clustering* analisou se (e como) a massa de dados de trabalho, uma vez subdividida por similaridade de acordo com as taxas citadas, permite a comparação entre municípios em diferentes meses, desde que suas taxas sejam comparáveis nos mesmos meses. A comparabilidade foi evidenciada durante a análise dos resultados da pesquisa, o que sugere que um modelo similar, uma vez melhor aferido por métricas confiáveis, serviria como uma interessante ferramenta preditiva para futuras flutuações (retrações) neste mercado profissional. Provavelmente também seria útil para enquadrar quaisquer outros municípios com taxas similares, em qualquer época, podendo ser usada em um esquema preditivo relativamente confiável. Ademais, permitiria a síntese ou planejamento de protocolos para decisões similares, em outras configurações compatíveis com as deste modelo.

A finalidade da clusterização seria o desenvolvimento de uma fonte de direcionamento para uma possível segmentação das decisões baseadas em dados, ou mais especificamente, baseada na separação entre (no caso, quatro) diferentes grupos, indicando abordagens que teriam efetividades similares para os locais atribuídos a cada grupo. Porém estas abordagens poderiam ser diferentes para cada grupo. No objeto de trabalho desta pesquisa, os diferentes municípios seriam dinamicamente enquadrados em um destes quatro grupos, conforme as taxas que apresentaram em um dado período, sejam estas simultaneamente baixas para as três taxas analisadas; intermediárias ou isoladamente maiores para uma ou outra delas (Gráf. 8).

Visivelmente os CSMs de cada *cluster*, bem como o CSM geral do processo, não foram altos, e estes valores de médios a baixos são uma característica desta massa de dados em particular, que é relativamente heterogênea e seus *clusters* são muito contíguos (praticamente justapostos, quando não sobrepostos em suas periferias). E, portanto, não foi possível obter *clusters* com CSMs muito altos (*i.e.*, com valores mais próximos de 1). A análise aqui demonstrada indicou, no entanto, que cada um destes 4 grupos identificados tem características muito distintas das dos demais (segundo as dimensões analisadas), sendo que:

a) um deles (o *Cluster*_2) foi muito coeso, e foi composto por municípios observados em meses que apresentaram todas as três taxas baixas. Este *cluster* foi o mais densamente populoso (de observações), com praticamente dois terços de todos os pontos;
b) o segundo (por ordem de número de componentes) foi o *Cluster*_0, que reuniu uma quinta parte dos pontos, que foram os municípios-mês com apenas as incidências mais altas;
c) o terceiro foi o *Cluster*_1, com cerca de um oitavo do total de pontos, que agregou os municípios-mês em que as incidências foram inferiores, porém com as taxas de hospitalização e de letalidade, embora não muito elevadas, relativamente maiores do que as incidências; e
d) o quarto foi o *Cluster*_3, com menos de dois centésimos do total de pontos, e que reuniu as taxas mais elevadas tanto de hospitalização quanto de letalidade, porém concentradas abaixo de um décimo das incidências.

A discussão acima foi baseada na Tab. 8, sobre a composição dos 4 *clusters*.

O modelo aqui desenvolvido (e otimizado) demonstrou como pode ser feito o processo de segmentação de uma massa de dados conforme determinados critérios de interesse, mensurados por variáveis específicas. A ferramenta pode tornar-se um importante aliado para a prática cotidiana de gestores que adotem variáveis relevantes para suas tomadas de decisão, após identificar grupos (*i.e.*, subgrupos) nos quais as variáveis de interesse assumem valores similares para cada subgrupo. No caso presente, a segmentação dos municípios (conforme taxas locais da COVID-19) pode facilitar o enquadramento de um dado local e a predição dos resultados avaliados pelas métricas correspondentes, como será discutido nas seções subsequentes. Portanto, parece ser muito útil ter-se de uma ferramenta que permita predizer circunstâncias futuras com base em eventos presentes ou passados, no próprio local ou em outros proximamente comparáveis a este (os que estiverem enquadrados no mesmo *cluster*).

5.3 RESPOSTAS À *SURVEY* E SEU ETL

Após ter sido feita a etapa de ETL dos resultados da *survey*, e cruzados seus dados com os dos agrupamentos por município-mês, chegou-se ao "material bruto" (*raw data*) para a aplicação dos algoritmos. Uma vez mais, esta etapa foi composta por uma sequência não condicional de operações simples.

O ETL mostrou ser uma tarefa mais prolongada, especificamente pelo tempo necessário para o treinamento do pesquisador no uso dos recursos do Knime, embora individualmente as tarefas propriamente ditas não tenham muita complexidade. Em poucas vezes recorreu-se ao uso do *node Column Expressions*, com a adoção do uso de comandos escritos em JavaScript. E, mesmo para estas poucas situações, o uso de códigos escritos em linguagem formal (*i.e.*, em

JavaScript), foi muito facilitado pelo material instrucional já visitado sobre a sintaxe de JavaScript e pelos recursos oferecidos no Knime.

O citado baixo número (22) de respostas válidas compromete a validade externa das conclusões da pesquisa, *i.e.*, a capacidade de generalização das conclusões da pesquisa para outros municípios, ou mesmo para retratar a realidade dos impactos econômicos da COVID-19 apenas para estes mesmos municípios. Na grande maioria dos municípios houve apenas uma resposta válida. Apesar da falta de representatividade das respostas coletadas, este volume de dados prestou-se à finalidade proposta da construção e teste de funcionalidade dos modelos gerados no Knime.

A distribuição ponderal (Seção 4.4) mostrou mais detalhadamente a falta de representatividade dos respondentes, frente ao total de dentistas do RS (e mesmo dentro dos próprios municípios), o que facilmente poderia explicar quaisquer dificuldades para a distinção entre as predições correspondentes às respostas coletadas e outras, feitas ao acaso (ou aleatoriamente). No entanto, este número de respostas, embora muito pequeno, serviu plenamente ao propósito de demonstrar como pode ser feito não somente o ETL, mas também todo o restante dos processos de ML, o desenvolvimento dos workflows, a seleção e a aplicação dos diversos algoritmos preditivos apresentados nesta pesquisa.

Uma das interpretações a que se pôde chegar por meio da EDA, relativamente às eventuais associações entre as taxas da COVID-19 e os fluxos nos consultórios (Tab. 8) foi a de que, caso fosse possível constatar uma associação entre qualquer das variáveis selecionadas para mensurar a COVID-19 e a das consultas efetivadas, a severidade parece ser a mais indicada para apresentar este tipo de associação.

5.4 ALGORITMOS DE REGRESSÃO

Os dados não se mostraram ajustados a qualquer dos cinco algoritmos de regressão testados, *viz.*, o de Regressão Linear, os três de Regressão Polinomial, ou o de Logística. Seja na avaliação dos *Scatter plots* ou das curvas ROC (Tab. 22) das métricas coletadas e das predições, nenhum dos algoritmos de regressão identificou associações entre as taxas e as métricas reais apuradas ou preditas, ao menos não muito melhores do que associações ao acaso.

Atribui-se a falta de adequação destes dados aos modelos de regressão principalmente à própria natureza da distribuição dos dados e à escassez de respostas à *survey*.

5.5 ALGORITMOS DE CLASSIFICAÇÃO

As classificações preditas pelos diferentes algoritmos de classificação selecionados, *viz*., o *k*-NN, o SVM, o Naïve Bayes, a NN dos MLP e a AutoML, se mostraram bastante diferentes entre si. Respectivamente, tiveram acurácias gerais (sob parâmetros otimizados) de pouco mais de um quarto para os dois primeiros; de pouco mais de três quartos para o terceiro, de quase a totalidade e a totalidade de acertos para os dois últimos. A partir das considerações anteriores decorrentes da limitação no número de respostas à *survey*, e da leitura das acurácias destes algoritmos (e por serem elas tão distintas entre si), permitem admitir que modelos gerados por workflows similares a estes, porém com maiores volumes de dados, possam ter performances melhores, e especialmente mais significativas. Um melhor treinamento de seus modelos pode lhes permitir atingir proporções mais altas de predições acertadas, cuja distribuição possa ser adequadamente descrita por seu agrupamento em classes (no caso, os seis intervalos de percentuais trabalhados ao longo da pesquisa).

Também pode ser destacado que as duas maiores acurácias, embora possam parecer-se com situações de *over-fitting*, não podem ser consideradas como tal, por causa da ausência de representatividade do número de respostas coletadas, e as próprias acurácias não deveriam *a priori*, ser tomadas como o produto de uma avaliação definitiva da situação pelo *app* gerado.

Ao contrário do que talvez possa parecer, a constatação acima não significa que, se as coletas de dados tivessem sido feitas através de variáveis contínuas (*e.g.*, como em uma *survey* com respostas fornecidas através de uma barra deslizante), os resultados não tivessem sido tão diferentes entre os dos algoritmos de classificação e os dos de regressão. Quando uma resposta (ou seu conjunto) deve ser enquadrada em uma "classe" (ou intervalo a que uma dada instância é atribuída, distinguindo-a das demais), diferentes outros valores próximos são reunidos em uma mesma classe. Assim, haverá uma proporção muito maior dos resultados que terão valores iguais à predição correspondente. Porém o nível de ajuste entre as predições de um modelo e as métricas dos valores reais correspondentes geralmente é maior (excetuando-se as predições acertadas ao acaso) para os algoritmos mais adequados a uma massa de dados específica. E isto torna ainda mais oportuno, ao desenvolver um projeto de ML, fazer-se alguns testes com mais de um algoritmo, para descobrir qual(ais) algoritmos preditivos – *e.g.*, os de regressão, ao se trabalhar com variáveis contínuas (dados numéricos) ou os de classificação, ao se tentar prever classes (ou intervalos discretos) em que os valores das variáveis devam ser enquadrados – se adaptam mais ao conjunto de dados de trabalho. De fato, existem centenas, milhares (ou mesmo

muitas vezes mais) de diferentes algoritmos, incluindo desde os mais clássicos e conhecidos até os desenvolvidos (e com boas performances) para *datasets* muito específicos, com pouca divulgação ou baixo número de pesquisas que os usem. E as características de uma massa de dados em particular é que determinará qual o algoritmo que melhor se adapta a ela e que gera predições com maior nível de acerto. Para os dados trabalhados – destacando-se as restrições, concessões ou comprometimentos nos resultados já comentados – foi observado que alguns dos algoritmos de classificação (especialmente em workflows já otimizados) tiveram performances marcadamente melhores do que as dos algoritmos de regressão.

No modelo aqui desenvolvido, foi intencionalmente escolhido um número de algoritmos, com o estrito fim de demonstrar como é feito o seu uso, sua comparação, e como este material pode ser usado como ponto de partida para construir workflows que incorporem diferentes algoritmos como ferramentas auxiliares para a execução de diversas tarefas de ML.

Os resultados da *survey* mostraram-se como o ponto mais frágil da pesquisa, porque, devido ao seu número marcadamente pequeno, podem simplesmente representar situações tão individuais que difiram demasiado das médias municipais (ou de qualquer outra medida de dispersão) (CALLEGARI-JACQUES, 2003). Para quase a totalidade dos municípios, as respostas se resumiram a um respondente por município. E estes casos individuais tanto poderiam ser *outliers* locais quanto os mais representativos de seus municípios. Ou mesmo qualquer combinação aleatória de efeitos (econômicos da COVID-19) distribuídos ao longo de um *continuum* entre ambos os extremos. Esta é a razão de ter sido enfatizado, ao longo da redação deste relatório, que os resultados (*i.e.*, os valores) obtidos podem não representar a realidade pesquisada, embora o processo de obtenção destes valores tenha demonstrado como pode ser efetivamente feita esta busca e predição, porém sendo que melhores performances seriam dependentes de melhor qualidade e quantidade dos dados de entrada.

Em todas as análises subsequentes, considerando as observações acima, destaca-se a lacuna de validade dos resultados numéricos obtidos e de eventuais conclusões que pudessem ser suscitadas durante a análise destes números, porém sim de seus padrões e parâmetros para os ajustes de funcionalidade da aplicação desenvolvida. Por outro lado, visando o objetivo principal da elaboração de um modelo funcional, verificável e flexível (*i.e.*, adaptável a outras questões e situações de pesquisa, e para outras massas coletadas de dados), o maior esforço de pesquisa foi centrado no próprio desenvolvimento e avaliação dos modelos, e na aplicação dos recursos que o Knime oferece para estas tarefas.

6 CONCLUSÕES

O relatório aqui apresentado mostrou que o Knime é uma das ferramentas de DS que permitem que pesquisadores e gestores investiguem diferentes massas de dados para identificar associações estatísticas entre variáveis e DBs aparentemente não relacionadas, instrumentalizando estes pesquisadores, sejam eles iniciantes ou não em DS, para um KDD e melhores tomadas de decisão baseadas em dados. Foi desenvolvido e testado, no Knime, um instrumento para a investigação do objeto específico de pesquisa, *viz*., os impactos econômicos da pandemia de COVID-19 sobre o mercado odontológico privado do RS, comparando a performance de 10 diferentes algoritmos de ML.

O principal objetivo da pesquisa não pôde ser plenamente beneficiado pelos recursos do Knime. Este objetivo geral não foi completamente atingido, em seus valores, devido à escassez de dados coletados na *survey*. Em contrapartida, foi devida e detalhadamente demonstrado todo o processo de construção de uma ferramenta (o *app*) para a investigação, identificação e quantificação destas associações e de suas probabilidades de acerto. Foram comparados dez diferentes algoritmos e suas performances relativas para a predição de valores com base nos treinamentos feitos e utilizando a validação cruzada. As performances relativas indicaram que os cinco algoritmos de regressão selecionados (a Linear, a Polinomial em graus 2, 3 e 4, e a Logística) não se mostraram adequados a esta massa de dados em particular, pois não geraram predições marcadamente mais acertadas do que as predições feitas ao acaso. Por outro lado, dentre os algoritmos de classificação selecionados, o *k*-NN e o SVM também se mostraram com uma performance muito limitada para estes dados específicos. E os demais algoritmos (Naïve Bayes, rede neural MLP e AutoML) tiveram performance superior a excelente com estes dados.

Alternativamente ao que foi executado, um teste e execução do mesmo *app*, porém com maiores massas de dados, demandaria mínimas adaptações (se alguma), e poderia revelar se estes níveis de desempenho se repetiriam com outros dados que representassem melhor esta mesma realidade investigada.

Constatou-se que o esforço de pesquisa para o desenvolvimento do *app*, em grande parte, independe do tamanho da massa de dados. A ferramenta apresentou grande capacidade de processamento, o que trouxe uma flexibilidade muito superior às expectativas para a pesquisa aqui descrita. Mesmo que não tivesse ocorrido a citada escassez de respostas à *survey*, *i.e.*, que houvesse um número grande de participantes (até mesmo em escalas censitárias), não teriam sido demandados esforços adicionais do desenvolvedor (*i.e.*, o pesquisador).

Por outro lado, cabe destacar que todo o desenvolvimento de *apps* – como foi demonstrado nos workflows ao longo do presente relatório, pela construção da ferramenta para a investigação, identificação e análise de possíveis associações estatísticas entre diferentes variáveis, presentes em DBs muito díspares entre si – foi feito mesmo sem a necessidade de um preparo ou treinamento prévio do pesquisador em linguagens formais de programação que exijam a digitação de linhas escritas de código. Esta restrição atendeu ao objetivo de demonstrar que outros profissionais não-programadores – entre eles, os da saúde (como os da Odontologia) – podem fazer pesquisas que usem esta ferramenta com interface gráfica (GUI), ou outras comparáveis, para aplicar técnicas da DS que permitam extrair conhecimento relevante sobre associações entre variáveis, sem, no entanto, depender de um longo e árduo investimento no seu próprio preparo e treinamento em linguagens formais de programação. Em outras palavras, foi atingido o objetivo de construção do aplicativo usando apenas recursos *low code* (ou mesmo, na quase totalidade do *app*, os recursos *no code*). Embora a consecução tenha demandado uma certa dose de empenho e persistência, não pareceu ser indispensável ter um treinamento formal em linguagens de programação para atingir este objetivo.

Uma parte importante do esforço de pesquisa, do tempo nela investido, e igualmente, do aprendizado no uso do Knime, foi direcionado à seleção das DBs que contivessem as variáveis mais vinculadas às grandezas (e das possíveis associações entre elas) que se desejava investigar, extração de seus valores, transformação de suas unidades e formatos, e o subsequente carregamento destes dados no Knime. Efetivamente, como previsto, esta fase absorveu a maior parte do período e empenho dedicados à pesquisa. Embora já comentado anteriormente que esta é uma constatação muito comum em projetos de DS, seria interessante ressaltar que este elemento deve ser computado durante a elaboração de um projeto de pesquisa, independentemente da ferramenta a ser escolhida para o estudo. Quaisquer das etapas da pesquisa, desde a seleção das DBs e de suas variáveis, passando pelo ETL, pela EDA, e até as tarefas de ML propriamente ditas – as quais permitem a identificação e o mapeamento de padrões e associações entre as variáveis, sejam estes imediatamente evidentes ou não – mostraram-se sequencialmente críticas para a consecução dos objetivos propostos.

Também foi observado que a baixa adesão dos colegas de classe à *survey*, fornecendo apenas um número muito limitado de respostas válidas, comprometeu principalmente a validade externa da pesquisa, porém não sua validade interna, no que tange à construção do *app* e à apresentação de uma forma de capacitação de eventuais futuros pesquisadores que decidam pela incorporação de recursos de DS às suas pesquisas. Devido a este pequeno número de

respostas, não foi possível atingir outro dos objetivos, o de mapear adequadamente as características dos profissionais que atuam na Odontologia privada do RS (embora os workflows apresentados indiquem como esta etapa da pesquisa poderia ter sido feita, caso fossem coletados dados relevantes e significativos através da *survey*). E esta tarefa permitiria identificar e caracterizar nichos mais suscetíveis ou mais resilientes aos fatores externos mensurados e que determinem em maior ou menor grau os efeitos investigados. Ou mesmo avaliar estas relações a outros fatores comparáveis. Portanto, não foi possível fazer o mapeamento das características dos profissionais da Odontologia privada do RS. Esta meta inicial (e que foi uma das norteadoras da pesquisa aqui apresentada) não foi atingida devido à grande escassez de respostas à *survey*, que determinou uma incapacidade discriminatória entre dados representativos e *outliers*, ou mesmo entre respostas próximas das medidas centrais de dispersão e respostas espúrias ou enviesadas.

Um achado relevante da pesquisa foi a constatação de que diferentes algoritmos podem ter performances preditivas marcadamente distintas, permitindo predizer, com níveis de probabilidade quantificados, qual (ou quais) podem ser os mais adequados para basear decisões a partir de massas específicas de dados. Um corolário desta conclusão é o de que parece ser conveniente que pesquisadores e gestores usem ferramentas que permitam comparar diferentes algoritmos e selecionar qual deles é o mais adequado para cada situação e cada massa de dados em particular.

Foi atingido o objetivo de demonstrar a construção de uma ferramenta para identificar as associações investigadas, embora a escassez de dados tenha inviabilizado a validade externa dos achados numéricos da pesquisa, o que permitiria projetar generalizações para todo o mercado odontológico privado regional, como se pretendera inicialmente.

REFERÊNCIAS BIBLIOGRÁFICAS

ADA, Subcommittee on Ethics. **Dental practice reopening: Following the American Dental Association Principles of Ethics and Code of Professional Conduct**. American Dental Association, 2020. Disponível em: https://doi.org/10.1016/j.adaj.2020.07.020. Acesso em: 23 nov. 2020.

AHA, David W. **Lazy Learning**. Dordrecht, Netherlands: Springer-Science+Business Media, 1997. Disponível em: https://doi.org/10.1007/978-94-017-2053-3. Acesso em: 10 nov. 2021.

AHUJA, Bhavdeep Singh. View of The Rising Costs in Oral Health Care Services - Time to Reboot & Reset Clinics for Economics Part – I. **World Journal of Advanced Scientific Research**, v. 3, n. 2, p. 43–63, 2020. Disponível em: https://wjasr.in/index.php/wjasr/article/view/44/34. Acesso em: 23 nov. 2020.

AMER, Faten *et al.* Assessment of Countries' Preparedness and Lockdown Effectiveness in Fighting COVID-19. **Disaster Medicine and Public Health Preparedness**, v. 15, n. 2, p. e15–e22, 2021. Disponível em: https://doi.org/10.1017/DMP.2020.217. Acesso em: 29 maio 2022.

ANDERSON, Roy M. *et al.* **How will country-based mitigation measures influence the course of the COVID-19 epidemic?**. Lancet Publishing Group, 2020. Disponível em: https://doi.org/10.1016/S0140-6736(20)30567-5. Acesso em: 23 nov. 2020.

BALDWIN, Richard; WEDER, Beatrice di Mauro. **Economics in the Time of COVID-19**. London: CEPR Press, 2020. *E-book.*

BASTANI, Peivand *et al.* Global concerns of dental and oral health workers during COVID-19 outbreak: a scope study on the concerns and the coping strategies. **Systematic Reviews**, v. 10, n. 45, 2021. Disponível em: https://doi.org/10.1186/s13643-020-01574-5. Acesso em: 02 fev. 2023.

BELLMAN, R. (Gerald Tesauro, David S. Touretzky, & Todd K. Leen, Org.)**NIPS'94: Proceedings of the 7th International Conference on Neural Information Processing Systems**. Denver, CO: MIT Press, 1994. p. 521–528. Acesso em: 12 nov. 2021.

BRENDER, J.; TALMON, J.; KEIZER, N.; NYKÄNEN, P.; RIGBY, M.; AMMENWERTH, E. STARE-HI - Statement on Reporting of Evaluation Studies in Health Informatics: explanation and elaboration. **Appl Clin Inform**, v. 4, n. 3, p. 331-58, 2013. Disponível em: https://doi: 10.4338/ACI-2013-04-RA-0024. Acesso em: 15 dez. 2022.

BÖGER, Beatriz *et al.* Systematic review with meta-analysis of the accuracy of diagnostic tests for COVID-19. **American Journal of Infection Control**, 2020. Disponível em: https://doi.org/10.1016/j.ajic.2020.07.011 Acesso em: 13 out. 2020.

BORGES, W. S. A Navalha de Ockham: Um Princípio Lógico de Parcimônia. **Scintilla – Revista de Filosofia e Mística Medieval**, v. 19, n. 1, p. 129–142, 2022. Disponível em: https://scintilla.saoboaventura.edu.br/scintilla/article/view/133. Acesso em: 19 dez. 2022.

BRASIL, Ministério da Saúde, Secretaria de Vigilância em Saúde. **Plano de preparação brasileiro para o enfrentamento de uma pandemia de influenza**. Brasília: MS, Editora,

2005. *E-book*.

CABRERA-TASAYCO, Fiorella Del Pilar *et al.* Biosafety measures at the dental office after the ap*p*earance of COVID-19: A systematic review. **Disaster Medicine and Public Health Preparedness**, 2020. Disponível em: https://doi.org/10.1017/dmp.2020.269. Acesso em: 17 nov. 2020.

CALLEGARI-JACQUES, Sidia Maria. **Bioestatística: Princípios e Aplicações**. Porto Alegre: Artmed, 2003.

CARRER, Fernanda Campos de Almeida *et al.* **Teleodontologia e SUS: uma importante ferramenta para a retomada da Atenção Primária à Saúde no contexto da pandemia de COVID-19**. Brasília: SciELO Preprints, 2020. Disponível em: https://doi.org/https://doi.org/10.1590/SciELOPreprints.837. Acesso em: 17 ago. 2020.

CAZELLA, Sílvio César; FEYH, Rafael; BEN, Ângela Jornada. A Decision Support System for Medical Mobile Devices Based on Clinical Guidelines for Tuberculosis. *In*: RAMOS, Carlos *et al.* (org.). **Ambient Intelligence - Software and Applications**. Cham, Switzerland: Springer International Publishing, 2014. p. 256. Disponível em: https://doi.org/10.1007/978-3-319-07596-9_24. Acesso em: 15 dez. 2022.

CDC, Centers for Disease Control and Prevention. **Guidance for dental settings**. 2020. Disponível em: https://www.cdc.gov/coronavirus/2019-ncov/hcp/dental-settings.html. Acesso em: 23 nov. 2020.

CHAMORRO-PETRONACCI, Cintia *et al.* Assessment of the Economic and Health-Care Impact of COVID-19 (SARS-CoV-2) on Public and Private Dental Surgeries in Spain: A Pilot Study. **International Journal of Environmental Research and Public Health**, v. 17, n. 14, p. 5139, 2020. Disponível em: https://doi.org/10.3390/ijerph17145139. Acesso em: 25 nov. 2020.

CHAN, Jasper Fuk Woo *et al.* A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: a study of a family cluster. **The Lancet**, v. 395, n. 10223, p. 514–523, 2020. Disponível em: https://doi.org/10.1016/S0140-6736(20)30154-9. Acesso em: 23 nov. 2020.

CHANDRA, Mukesh *et al.* Digital technologies, healthcare and Covid-19: insights from developing and emerging nations. **Health and Technology**, v. 12, n. 2, p. 547–568, 2022. Disponível em: https://doi.org/10.1007/s12553-022-00650-1. Acesso em: 15 dez. 2022.

CHECCHI, Vittorio *et al.* COVID-19 dentistry-related aspects: a literature overview. **International Dental Journal**, p. idj.12601, 2020. Disponível em: https://doi.org/10.1111/idj.12601. Acesso em: 23 nov. 2020.

CONSELHO NACIONAL DE SAÚDE. **Resolução n° 466**. 2013. Disponível em: http://conselho.saude.gov.br/resolucoes/reso_12.htm. Acesso em: 23 mar. 2020.

CORTES, Carinna; VAPNIK, Vladimir. Support Vector Networks. **Machine Leaming**, v. 20, p. 279–297, 1995. Disponível em: https://doi.org/10.1007/BF00994018 Acesso em: 18 out. 2020.

COSTA, Ligia Maria Cantarino da; MERCHAN-HAMANN, Edgar. Pandemias de influenza

e a estrutura sanitária brasileira: breve histórico e caracterização dos cenários. **Revista Pan-Amazônica de Saúde**, v. 7, n. 1, p. 11–25, 2016. Disponível em: https://doi.org/10.5123/s2176-62232016000100002. Acesso em: 10 jul. 2021.

COTRIN, Paula *et al.* Impact of coronavirus pandemic in appointments and anxiety/concerns of patients regarding orthodontic treatment. **Orthodontics and Craniofacial Research**, v. 23, n. 4, p. 455–461, 2020. Disponível em: https://doi.org/10.1111/ocr.12395. Acesso em: 9 jul. 2021.

CRISTIANINI, Nello; SHAWE-TAYLOR, John. **An Introduction to Support Vector Machines and Other Kernel-based Learning Methods**. Cambridge: Cambridge University Press, 2014. Disponível em: https://doi.org/10.1017/CBO9780511801389. Acesso em: 25 ago. 2022.

CVETKOVIĆ, Vladimir M. *et al.* A Predictive Model of Pandemic Disaster Fear Caused by Coronavirus (COVID-19): Implications for Decision-Makers. **International Journal of Environmental Research and Public Health 2022**, v. 19, p. 652, v. 19, n. 2, p. 652, 2022. Disponível em: https://doi.org/10.3390/IJERPH19020652. Acesso em: 29 jan. 2023.

DA CUNHA, Amanda Ramos *et al.* The impact of the COVID-19 pandemic on the provision of dental procedures performed by the Brazilian Unified Health System: a syndemic perspective. **Revista Brasileira de Epidemiologia**, v. 24, n. E210028, p. 1–10, 2021. Disponível em: https://doi.org/10.1590/1980-549720210028. Acesso em: 27 abr. 2022.

DE MELO, Márcio Cristiano *et al.* Uma análise bibliométrica das pesquisas globais da COVID-19. **InterAmerican Journal of Medicine and Health**, v. 3, 2020. Disponível em: https://doi.org/10.31005/iajmh.v3i0.88. Acesso em: 17 ago. 2020.

DEL RIO, Carlos; MALANI, Preeti N. **COVID-19 - New Insights on a Rapidly Changing Epidemic**. American Medical Association, 2020. Disponível em: https://doi.org/10.1001/jama.2020.3072. Acesso em: 23 nov. 2020.

FAYYAD, Usama; PIATETSKY-SHAPIRO, Gregory; SMYTH, Padhraic. From Data Mining to Knowledge Discovery in Databases. **AI Magazine**, v. 17, n. 3, p. 37–37, 1996. Disponível em: https://doi.org/10.1609/AIMAG.V17I3.1230. Acesso em: 8 jul. 2022.

FIX, Evelyn; HODGES JR., J. L. **Discriminatory Analysis Nonparametric discrimination:** Consistency properties. Randolph Field, Tx: 1951.

FLANAGAN, David. **JavaScript The Defini Guide**. 6. ed. Sebastopol, CA: O'Reilly Media, 2011.

FREITAS, Henrique *et al.* O método de pesquisa survey. **RAUSP - Revb. Administração da USP**, v. 35, n. 3, 2000. Disponível em: http://www.ufrgs.br/gianti/files/artigos/2000/2000_092_RAUSP.PDF. Acesso em: 26 nov. 2020.

GAO, George F. **From "A"IV to "Z"IKV: Attacks from Emerging and Re-emerging Pathogens**. Cell Press, 2018. Disponível em: https://doi.org/10.1016/j.cell.2018.02.025. Acesso em: 8 jan. 2022.

GARCÉS-ELÍAS, María Claudia *et al.* Impact of mandatory social isolation measures due to

the COVID-19 pandemic on the subjective well-being of Latin American and Caribbean dentists. **Journal of Clinical and Experimental Dentistry**, v. 14, n. 1, p. e40, 2022. Disponível em: https://doi.org/10.4317/JCED.58776. Acesso em: 9 fev. 2023.

GRAY, Richard; SANDERS, Chris. A reflection on the impact of COVID-19 on primary care in the United Kingdom. **Journal of interprofessional care**, v. 34, n. 5, p. 672–678, 2020. Disponível em: https://doi.org/10.1080/13561820.2020.1823948. Acesso em: 8 jul. 2021.

GRECO, Christohper. **Data Science Tools - R, Excel, KNIME & OpenOffice**. Dulles, VA: Mercury Learning and Informatnion, 2020. *E-book*.

GUAN, W. *et al.* Clinical characteristics of coronavirus disease 2019 in China. **New England Journal of Medicine**, v. 382, n. 18, p. 1708–1720, 2020. Disponível em: https://doi.org/10.1056/NEJMoa2002032. Acesso em: 3 nov. 2020.

HARO, Juan Carlos de *et al.* Psychological Impact of COVID-19 in the Setting of Dentistry: A Review Article. **International Journal of Environmental Research and Public Health**, v. 19, n. 16216, p. 1–37, 2022. Disponível em: https://doi.org/10.3390/ijerph192316216. Acesso em: 8 nov. 2022.

HINKEL, José Henrique Schwanck. **Processo de Data Linkage para qualificação das bases de notificação de Covid-19: Um Estudo de Caso**. 149 f. 2022. - Universidade Federal de Ciências da Saúde de Porto Alegre, 2022. Disponível em: https://repositorio.ufcspa.edu.br/jspui/handle/123456789/1934. Acesso em: 15 dez. 2022.

HUGO, F.N.; KASSEBAUM, N.J.; BERNABÉ, E. Role of Dentistry in Global Health: Challenges and Research Priorities. **Journal of Dental Research**, v. 100, n. 7, 2021. Disponível em: https://doi.org/https://doi.org/10.1177/0022034521992011. Acesso em: 27 abr. 2022.

IHME, Institute for Health Metrics and Evaluations. **IHME COVID-19 Projections**. 2022. Disponível em: https://covid19.healthdata.org/global?view=daily-deaths&tab=trend. Acesso em: 25 mar. 2022.

IVAKHNENKO, Alexey Grigoryevich; LAPA, Valentin Grigoryevich. **Cybernetic Predicting Devices**. Washington, D.C.: 1965. Disponível em: https://www.gwern.net/docs/ai/1966-ivakhnenko.pdf. Acesso em: 27 abr. 2022.

IYENGAR, Karthikeyan *et al.* Learning opportunities from COVID-19 and future effects on health care system. **Diabetes and Metabolic Syndrome: Clinical Research and Reviews**, v. 14, n. 5, p. 943–946, 2020. Disponível em: https://doi.org/10.1016/j.dsx.2020.06.036. Acesso em: 18 abr. 2022.

KEERTHI, S. S. *et al.* Improvements to Platt's SMO Algorithm for SVM Classifier Design. **Neural Computation**, Singapore, v. 13, n. 3, p. 637–649, 2001. Disponível em: https://doi.org/10.1162/089976601300014493. Acesso em: 13 abr. 2022.

KETCHEN JR., David J.; SHOOK, Christopher L. The application of cluster analysis in stretegic management research: An analysis and critique. **Strategic Management Journal**, v. 17, n. 1, p. 441–458, 1996. Disponível em: https://onlinelibrary-wiley.ez45.periodicos.capes.gov.br/doi/epdf/10.1002/%28SICI%291097-0266%28199606%2917%3A6%3C441%3A%3AAID-SMJ819%3E3.0.CO%3B2-G. Acesso

em: 25 ago. 2022.

KRZANOWSKI, Wojtek J.; HAND, David J. **ROC curves for continuous data**. Boca Raton, FL: CRC Press, 2009.

LANA, Raquel Martins *et al.* Emergência do novo coronavírus (SARS-CoV-2) e o papel de uma vigilância nacional em saúde oportuna e efetiva. **Cadernos de Saúde Pública**, v. 36, n. 3, 2020. Disponível em: https://doi.org/10.1590/0102-311X00019620. Acesso em: 3 nov. 2022.

LI, Qun *et al.* Early transmission dynamics in Wuhan, China, of novel coronavirus-infected pneumonia. **N Engl J Med**, v. 382, n. 13, p. 1199–1207, 2020. Disponível em: https://doi.org/10.1056/NEJMoa2001316. Acesso em: 24 mar. 2020.

LINNAINMAA, Seppo. Taylor expansion of the accumulated rounding error. **BIT**, v. 16, n. 2, p. 146–160, 1976. Disponível em: https://doi.org/10.1007/BF01931367. Acesso em: 25 ago. 2022.

MACÁRIO, Carla Geovana do N.; BALDO, Stefano Monteiro. **O Modelo Relacional**. Campinas, SP: 2005. Disponível em: https://www.ic.unicamp.br/~geovane/mo410-091/Ch03-RM-Resumo.pdf. Acesso em: 25 ago. 2022.

MACQUEEN, J. Some methods for classification and analysis of multivariate observations. *In*: , 1967, Los Angeles. **Proceedings of the Berkeley symposium on mathematical statistics and probability**. Los Angeles: 1967. p. 281–297. Disponível em: https://projecteuclid.org/ebooks/berkeley-symposium-on-mathematical-statistics-and-probability/Some-methods-for-classification-and-analysis-of-multivariate-observations/chapter/Some-methods-for-classification-and-analysis-of-multivariate-observations/bsms. Acesso em: 3 jun. 2022.

MADHULATHA, T. Soni. An Overview on Clustering Methods. **IOSR Journal of Engineering**, v. 2, n. 4, p. 719–725, 2012. Disponível em: https://www.academia.edu/12394768/an_overview_on_clustering_methods. Acesso em: 12 ago. 2022.

MAI, Scheila *et al.* O uso das Tecnologias na Democratização da Informação em Saúde. **Revista de Gestão em Sistemas de Saúde -RGSS**, v. 6, n. 3, p. 210–218, 2017. Disponível em: https://doi.org/10.5585/rgss.v6i3.287. Acesso em: 25 nov. 2022.

MCCULLOCH, Warren S.; PITTS, Walter. A logical calculus of the ideas immanent in nervous activity. **The Bulletin of Mathematical Biophysics**, v. 5, n. 4, p. 115–133, 1943. Disponível em: https://doi.org/10.1007/BF02478259. Acesso em: 6 dez. 2022.

MENG, L.; HUA, F.; BIAN, Z. Coronavirus Disease 2019 (COVID-19): Emerging and Future Challenges for Dental and Oral Medicine. **Journal of Dental Research**, v. 99, n. 5, p. 481–487, 2020. Disponível em: https://doi.org/10.1177/0022034520914246. Acesso em: 8 jul. 2021.

MOHER, David *et al.* **Guidelines for reporting health research**: a user's manual. Wiley-Blackwell, 2014. *E-book.*

N., Pradeep; KAUTISH, Sandeep; PENG, Sheng-Lung. **Demystifying Big Data, Machine**

Learning, and Deep Learning for Healthcare Analytics. Cambridge, MA: Academic Press, Elsevier, 2021. Disponível em: https://doi.org/10.1016/B978-0-12-821633-0.00008-8. . Acesso em: 8 jan. 2023.

OLIVAL, Kevin J.; WEEKLEY, Cristin C.; DASZAK, Peter. Are Bats Really "Special" as Viral Reservoirs? What We Know and Need to Know. *In*: **Bats and Viruses: A New Frontier of Emerging Infectious Diseases**. Wiley, 2015. p. 281–294. Disponível em: https://doi.org/10.1002/9781118818824.ch11. Acesso em: 23 nov. 2020.

OPAS, Organização Panamericana para a Saúde. **Módulos de Principios de Epidemiología para el Control de Enfermedades**. Brasília: OPAS, Organização Panamericana da Saúde - Representação Basil, 2010. *E-book*.

PATEL, Naiya. Impact on Dental Economics and Dental Healthcare Utilization in COVID-19: An Exploratory Study. **Journal of Advanced Oral Research**, v. 11, n. 2, p. 128–136, 2020. Disponível em: https://doi.org/10.1177/2320206820941365. Acesso em: 23 nov. 2020.

PEARSON, Karl. VII. Note on regression and inheritance in the case of two parents. **Proceedings of the Royal Society of London**, v. 58, n. 347–352, p. 240–242, 1895. Disponível em: https://doi.org/10.1098/RSPL.1895.0041. Acesso em: 1 fev. 2023.

PELOSO, R. M. *et al.* How does the quarantine resulting from COVID-19 impact dental appointments and patient anxiety levels? **Brazilian Oral Research**, v. 34, 2020. Disponível em: https://doi.org/10.1590/1807-3107BOR-2020.VOL34.0084. Acesso em: 8 jul. 2021.

PEREIRA, Luciano José *et al.* Biological and social aspects of Coronavirus Disease 2019 (COVID-19) related to oral health. **Brazilian Oral Research**, v. 34, 2020. Disponível em: https://doi.org/10.1590/1807-3107bor-2020.vol34.0041. Acesso em: 13 jul. 2021.

PLATT, John C. **Fast Training of Support Vector Machines using Sequential Minimal Optimization**. Redmond, WA: 2000. Disponível em: https://www.microsoft.com/en-us/research/publication/fast-training-of-support-vector-machines-using-sequential-minimal-optimization/. Acesso em: 7 set. 2021.

POTTER, C. W. A history of influenza. *In*: , 2001. **Journal of Applied Microbiology**. 2001. p. 572–579. Disponível em: https://doi.org/10.1046/j.1365-2672.2001.01492.x. Acesso em: 10 jul. 2021.

QUISPE-JULI, Cender *et al.* COVID-19: Una pandemia en la era de la salud digital. **Unidad Inform Biomed Salud Glob.**, p. 1–19, 2020. Disponível em: https://doi.org/10.1590/SCIELOPREPRINTS.164. Acesso em: 11 jul. 2021.

RANJAN, Rohit; AGARWAL, Swati; VENKATESAN, Dr. S. Detailed Analysis of Data Mining Tools. **International Journal of Engineering Research & Technology**, v. 6, n. 5, p. 785–789, 2017. Disponível em: https://doi.org/10.17577/IJERTV6IS050459. Acesso em: 13 jun. 2022.

RICHTERICH, Peter. **Severe underestimation of COVID-19 case numbers: Effect of epidemic growth rate and test restrictionsmedRxiv**. 2020. Disponível em: https://doi.org/10.1101/2020.04.13.20064220. Acesso em: 5 out. 2020.

RIEDMILLER, Martin; BRAUN, Heinrich. **A Direct Adaptive Method for Faster**

Backpropagation Learning: The RPROP Algorithm. 1993. Disponível em: https://doi.org/10.1109/ICNN.1993.298623. Acesso em: 13 nov. 2021.

ROUSSEEUW, P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. **Journal of Computational and Applied Mathematics**, v. 20, n. C, p. 53–65, 1987. Disponível em: https://doi.org/10.1016/0377-0427(87)90125-7. Acesso em: 11 jun. 2022.

RUSSEL, S.; NORVIG, P. **Inteligência artificial**. 3. ed. Rio de Janeiro: Elsevier, 2013.

SANTOS, Alessandro S *et al.* Sistema de Monitoramento Inteligente da COVID-19 em SP. **Anais Estendidos do Simpósio Brasileiro de Sistemas Multimídia e Web (WebMedia)**, p. 87–90, 2021. Disponível em: https://doi.org/10.5753/WEBMEDIA_ESTENDIDO.2021.17619. Acesso em: 9 fev. 2023.

SCHWENDICKE, F.; SAMEK, W.; KROIS, J. Artificial Intelligence in Dentistry: Chances and Challenges: **https://doi.org/10.1177/0022034520915714**, v. 99, n. 7, p. 769–774, 2020. Disponível em: https://doi.org/10.1177/0022034520915714. Acesso em: 24 ago. 2021.

SILVEIRA, Anny Jackeline Torres. A medicina e a influenza espanhola de 1918. **Tempo**, v. 10, n. 19, p. 91–105, 2005. Disponível em: https://doi.org/10.1590/s1413-77042005000200007. Acesso em: 10 dez. 2021.

SOARES, Paulo Roberto Rodrigues *et al.* **A Pandemia de COVID-19 no RS e na metrópole de Porto Alegre**. Porto Alegre: 2020. Disponível em: https://www.observatoriodasmetropoles.net.br/wp-content/uploads/2020/07/Dossiê-Núcleo-Porto-Alegre_Análise-Local_Julho-2020.pdf. Acesso em: 12 nov. 2022.

SPEARMAN, C. The Proof and Measurement of Association between Two Things. **The American Journal of Psychology**, v. 15, n. 1, p. 72–101, 1904. Disponível em: https://doi.org/10.2307/1412159. Acesso em: 1 fev. 2023.

SPECHT, Donald F. Probabilistic neural networks. **Neural Networks**, v. 3, n. 1, p. 109–118, 1990. Disponível em: https://doi.org/10.1016/0893-6080(90)90049-Q. Acesso em: 13 nov. 2021.

THORNDIKE, Robert. Who belongs in the family? **Psychometrika**, v. 18, n. 4, p. 267–276, 1953. Disponível em: https://link.springer.com/content/pdf/10.1007/BF02289263.pdf?pdf=button. Acesso em: 11 jun. 2022.

TOBLER, W. R. A Computer Movie Simulating Urban Growth in the Detroit Region. **Economic Geography**, v. 46, n. Jun., p. 234–240, 1970.

TUKEY, John W. **Exploratory Data Analysis**. Mento Park, California: Addison-Wesley, 1977.

UJVARI, Stefan Cunha. **A História e suas Epidemias** - a convivência do Homem com os microorganismos. 2. ed. São Paulo: São Paulo: Senac São Paulo; Rio de Janeiro: Senac Rio, 2003. *E-book.*

UJVARI, Stefan Cunha. **Pandemias**: a humanidade em risco. São Paulo: Contexto, 2011.

VENKATESWARAN, Jayendran; DAMANI, Om. **Effectiveness of Testing, Tracing, Social Distancing and Hygiene in Tackling Covid-19 in India: A System Dynamics Model**. 2020. Acesso em: 8 jul. 2022.

VENKATESWARLU, P.; SPANADNA, R R.; SRIKANTH, K. An Extensive Study of Data Analysis Tools (Rapid Miner, Weka, R Tool, Knime, Orange). **SSRG International Journal of Computer Science and Engineering**, v. 5, n. 9, p. 4–11, 2018. Disponível em: https://doi.org/10.14445/23488387/IJCSE-V5I9P102. Acesso em: 12 jul. 2022.

WESTON, J.; WATKINS, C. **Multi-class Support Vector Machines**. Royal Holloway - University of London: Surrey, England: 1998. Acesso em: 17 ago. 2022.

WHO. **Critical Preparedness, Readiness and Response Actions for COVID-19**: Interim Guidance. World Health Organization. 2020. Disponível em: https://www.who.int/publications/i/item/critical-preparedness-readiness-and-response-actions-for-covid-19. Acesso em: 23 nov. 2020.

WORLD HEALTH ORGANIZATION, WHO. **WHO Director-General's opening remarks at the media briefing on COVID-19 - 11 March 2020**. 2020. Disponível em: https://www.who.int/dg/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020. Acesso em: 25 mar. 2020.

WU, Zunyou; MCGOOGAN, Jennifer M. **Characteristics of and Important Lessons from the Coronavirus Disease 2019 (COVID-19) Outbreak in China**: Summary of a Report of 72314 Cases from the Chinese Center for Disease Control and Prevention. American Medical Association, 2020. Disponível em: https://doi.org/10.1001/jama.2020.2648. Acesso em: 23 nov. 2020.

ZAMBON, Maria. **Influenza and other emerging respiratory viruses**. 2014. Disponível em: https://doi.org/10.1016/j.mpmed.2013.10.017. Acesso em: 10 jul. 2021.

ZHANG, Jin. **Visualization for Information Retrieval**. Milwaukee, WI: Springer-Verlag Berlin, 2008.

ZHU, Na *et al.* A novel coronavirus from patients with pneumonia in China, 2019. **The New England Journal of Medicine**, v. 382, n. 8, p. 727–733, 2020. Disponível em: https://doi.org/10.1056/NEJMoa2001017. Acesso em: 24 mar. 2020.

ZIMMERLING, Amanda; CHEN, Xiongbiao. Innovation and possible long-term impact driven by COVID-19: Manufacturing, personal protective equipment and digital technologies. **Technology in Society**, v. 65, p. 101541, 2021. Disponível em: https://doi.org/10.1016/J.TECHSOC.2021.101541

APÊNDICE A – Carta-convite de participação na pesquisa

Prezado(a) [nome do destinatário do contato via eletrônica],

Por meio desta, você está sendo convidado a participar da pesquisa "Efeitos da pandemia de COVID-19 no mercado privado da odontologia no estado do RS: Uma *survey* com os responsáveis por consultórios e clínicas", que avaliará os impactos econômicos da pandemia de COVID-19 no atual mercado privado da Odontologia no RS.

Para participar da pesquisa, você deve preencher, cumulativamente, os seguintes requisitos:

a) Atuar, exclusiva ou parcialmente, em gestão na Odontologia privada;

b) Exercer atividades de gestão ou relacionadas, como contabilidade, estoques, fluxo de caixa, etc., no consultório ou clínica odontológica, seja na administração de todo o estabelecimento, seja na condução de carteira própria de clientes/pacientes;

c) Os participantes não devem ter quaisquer restrições ou conflitos de interesse com a presente pesquisa, e declararão ter plena ciência de que é uma pesquisa estritamente acadêmica, sem finalidades comerciais, publicitárias ou financeiras.; e

d) Ter seus rendimentos vinculados à produtividade ou rentabilidade do consultório, *i.e.*, não receberem exclusivamente rendimentos fixos a intervalos regulares (*e.g.*, salários mensais).

O presente estudo será observacional, de cunho descritivo e empregará uma metodologia de natureza quantitativa do tipo *survey*, com a duração estimada de 15 a 30 min.

A pesquisa usará um questionário online na plataforma *Google Forms*, e o acesso às questões só é habilitado aos participantes após sua leitura e concordância com um Termo de Consentimento Livre e Esclarecido (TCLE), disponibilizado no início do questionário. É recomendado que os participantes guardem uma via destes documentos eletrônicos.

A pesquisa não tem qualquer risco ou desconforto físico, mas pode gerar algum desconforto psicológico ou emocional, por tratar de temas econômicos com impacto sobre a qualidade de vida e o exercício profissional dos participantes.

Todos os participantes declararão ter plena ciência de não receber benefícios diretos por tomar parte nela, de qualquer natureza e montante, porém apenas indiretos, por contribuírem para os objetivos desta pesquisa. Não há custos pela sua participação na pesquisa. Quaisquer custos diretos ou indiretos que você tiver por participar dela lhe serão ressarcidos pelo projeto da pesquisa, desde que previamente combinados entre pesquisadores e participantes.

Aos participantes e seus estabelecimentos é garantido total anonimato, uma vez que não fornecerão dados que permitam identificação dos participantes e/ou de seus estabelecimentos. As questões visam caracterizar o perfil de formação e atuação dos profissionais e dos estabelecimentos privados do mercado-alvo, bem como dos efeitos (mensurados em variação percentual) que a pandemia lhes trouxe.

O protocolo foi aprovado com o CAAE: 47666021.2.0000.5347, pelo Comitê de Ética em Pesquisa da UFRGS (Sistema CEP/CONEP), que é um órgão colegiado, de caráter consultivo, deliberativo e educativo, cuja finalidade é avaliar, emitir parecer e acompanhar os projetos de pesquisa envolvendo seres humanos, em seus aspectos éticos e metodológicos, realizados na instituição.

Todos os participantes terão plena liberdade para a interrupção ou cancelamento do preenchimento do questionário, antes de seu envio, no caso de ocorrência de qualquer eventual desconforto durante sua realização. Em tais casos, todos os dados até então coletados serão total e definitivamente eliminados da pesquisa. Após o envio, os dados coletados anonimamente integrarão um banco de dados no qual não será possível sua identificação, edição ou exclusão. Se você tiver dúvidas adicionais, pode contatar: a) Comitê de Ética em Pesquisa da UFRGS, F: (51) 3308-3738, e-mail etica@propesq.ufrgs.br; endereço: Av. Paulo Gama, 110, s. 311 - Anexo I da Reitoria - Campus Centro - Porto Alegre/RS; b) o pesquisador principal, Prof. Dr. Fernando Neves Hugo / fernandoneveshugo@gmail.com; ou c) Faculdade de Odontologia da UFRGS - R. Ramiro Barcelos, 2492 - Santa Cecilia, Porto Alegre/RS, F: (51) 3308-5010. Durante a pandemia de COVID-19, estes contatos devem ser feitos apenas por e-mail.

O link de acesso ao questionário é: < https://forms.gle/dT4pED1wMrTHdKdFA >.

Agradecemos antecipadamente por sua contribuição para uma melhor compreensão sobre este tema, que atinge tão diretamente a tantos colegas que atuam no mercado privado.

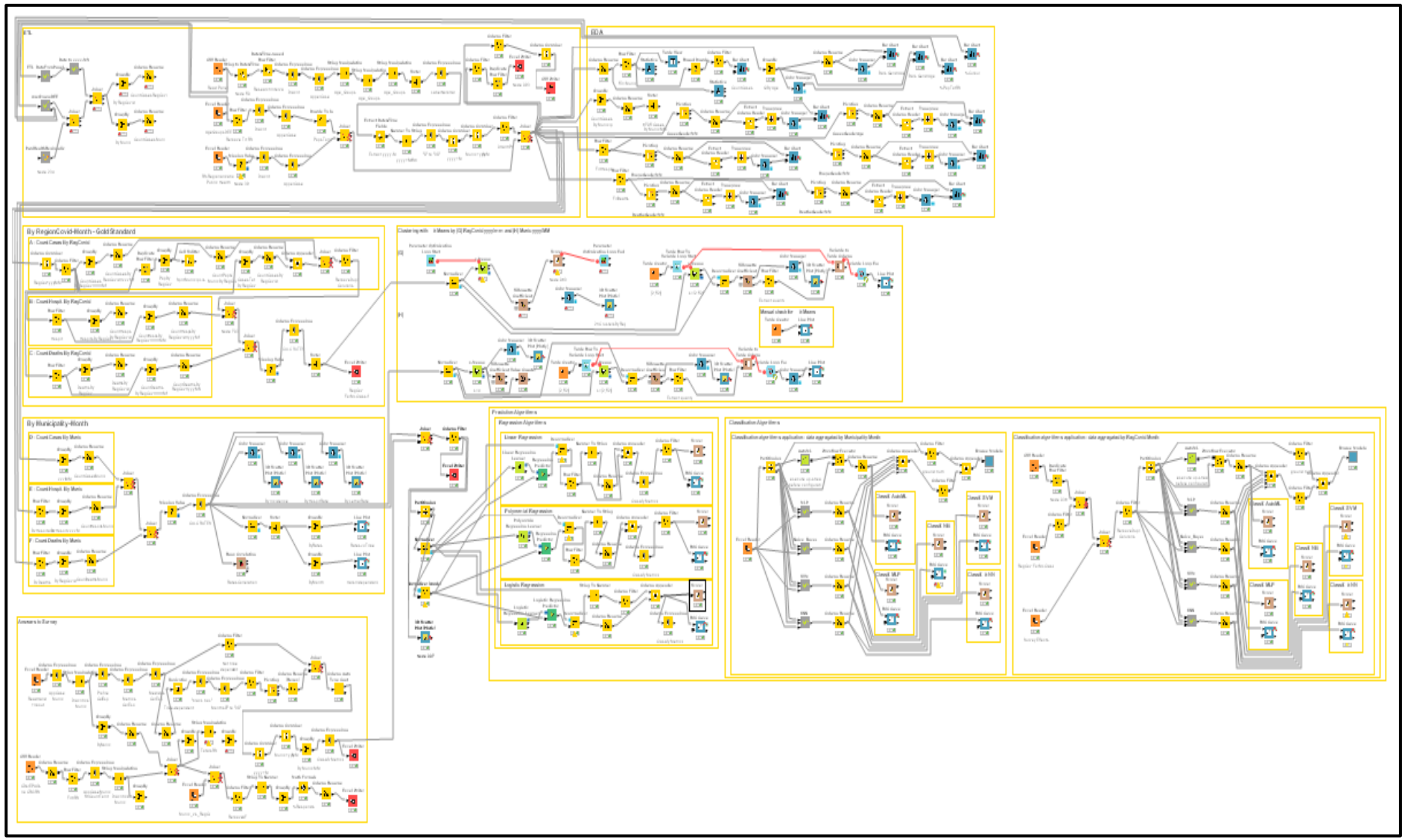
Atenciosamente. Os pesquisadores.

APÊNDICE B – Rótulos para variáveis da *survey*

Variável categórica	Valores originais	Valores pós-transformação
As “supostas” variáveis independentes (classificadores ou preditores)		
Ano de nascimento	já era numérica	Iguais
Ano da 1ª formatura	já era numérica	Iguais
Gênero	Femin.; Masc.; NBin;PrefNDiz;	0; 1; 2; 3
Formação	ñCD; CDsemPG; CDCapacAperf; CDEspec; CDMestr; CDEspecMest; CDDout; CDEspMesDout	0; 1; 2; 3; 4; 5; 6; 7
Experiência PublPriv	Nunca; NuncaPrivSemprePubl; ExPrivAtualSoPubl; AtualAmbas; ExPublAtualSoPriv; NuncaPublSemprePriv	0; 1; 2; 3; 4; 5; 6
Rotina Gestão ou Clínica	ÑclinÑgest ; SoClinNuncaGest; SoClinExGest; GestÑCD; ClinEGest	0; 1; 2; 3; 4
TempoGestãoPriv	Nunca; 0-1 ano; 1-5 anos; 5-10 anos; 10-20 anos; >20 anos	0; 1; 2; 3; 4; 5
TempoAssalarPriv	Nunca; 0-1 ano; 1-5 anos; 5-10 anos; 10-20 anos; >20 anos	0; 1; 2; 3; 4; 5
TempoRemunHoraTurno	Nunca; 0-1 ano; 1-5 anos; 5-10 anos; 10-20 anos; >20 anos	0; 1; 2; 3; 4; 5
TempoLocTurnos	Nunca; 0-1 ano; 1-5 anos; 5-10 anos; 10-20 anos; >20 anos	0; 1; 2; 3; 4; 5
TempoConsultPropr	Nunca; 0-1 ano; 1-5 anos; 5-10 anos; 10-20 anos; >20 anos	0; 1; 2; 3; 4; 5
TempoGestÑCDConsultPropr	Nunca; 0-1 ano; 1-5 anos; 5-10 anos; 10-20 anos; >20 anos	0; 1; 2; 3; 4; 5
ÉCDGestOUConsultPropr	NãoTemConsultPropr; CDContrataGestor; CDGestEAdmin; PrimAdminApósCD; CDGestÑAdmin; GestAdminÑCD; GestorÑAdminÑCD	0; 1; 2; 3; 4; 5; 6
TempoSemGestão	SempreEmGestão; 0-1 ano; 1-5 anos; 5-10 anos; 10-20 anos; NuncaFezGestão	0; 1; 2; 3; 4; 5
GestãoDuranteCOVID-19	ÑCoincGestCOV; InicGestDuranteAtualÑGest; GestAntesAtualÑGest; InicGestDuranteESemInterrup; DesdeAntesSemInterrup	0; 1; 2; 3; 4
N°MesesGestCOVID-19	já era numérica	iguais
CEPResp	Manter categórica	Iguais
LocalConsultNaCidade	Rural; Afast; Central	0; 1; 2
*** EspecialidAtendidas	***	***
N°Cadeiras	já era numérica	Iguais
ModeloRemuner	SoConven; SoCooper; Misto; SoPartic;	0; 1; 2; 3
%Partic	já era numérica (%)	Iguais
%Conven	já era numérica (%)	Iguais
%Cooper	já era numérica (%)	Iguais
As “supostas” métricas, ou 8 variáveis (supostamente) dependentes das anteriores		
FuncionamentoPorBimestre	Normal; <VolPac; SoUrgenc; Fechad	0; 1; 2; 3
Var%DemandaPorConsultas	< 5%; 5-10%; 11-24%; 25-50%; 51-75%; >75%	0,05; 0,1; 0,2; 0,4; 0,6; 0,8
Var%ConsultEfetiv	< 5%; 5-10%; 11-24%; 25-50%; 51-75%; >75%	0,05; 0,1; 0,2; 0,4; 0,6; 0,8
Var%TempoOcupCadeiras	< 5%; 5-10%; 11-24%; 25-50%; 51-75%; >75%	0,05; 0,1; 0,2; 0,4; 0,6; 0,8
Var%CustosFixos	< 5%; 5-10%; 11-24%; 25-50%; 51-75%; >75%	0,05; 0,1; 0,2; 0,4; 0,6; 0,8
Var%CustosVariáveis	< 5%; 5-10%; 11-24%; 25-50%; 51-75%; >75%	0,05; 0,1; 0,2; 0,4; 0,6; 0,8
Var%FaturBrut	< 5%; 5-10%; 11-24%; 25-50%; 51-75%; >75%	0,05; 0,1; 0,2; 0,4; 0,6; 0,8
Var%Lucrativ	< 5%; 5-10%; 11-24%; 25-50%; 51-75%; >75%	0,05; 0,1; 0,2; 0,4; 0,6; 0,8
Outras métricas (numéricas/booleanas) para variações no pessoal e modelo de negócio		
N°CDs	já era numérica (%)	Iguais
N°CDsProprSoc	já era numérica (%)	Iguais
%CDsProprSocSuspTotRetir	já era numérica (%)	Iguais
%CDsProprSocSuspParcRetir	já era numérica (%)	Iguais

%CDsProprSocSemVarRetir	já era numérica (%)	Iguais
N°CDsLocat	já era numérica (%)	Iguais
%CDsLocatSuspTotRetir	já era numérica (%)	Iguais
%CDsLocatSuspParcRetir	já era numérica (%)	Iguais
%CDsLocatSemVarRetir	já era numérica (%)	Iguais
N°CDsLocatPrestad	já era numérica (%)	Iguais
%CDsLocatSuspTotRetir	já era numérica (%)	Iguais
%CDsLocatSuspParcRetir	já era numérica (%)	Iguais
%CDsLocatSemVarRetir	já era numérica (%)	Iguais
N°CDsAssalar	já era numérica (%)	Iguais
%CDsAssalarSuspTotRetir	já era numérica (%)	Iguais
%CDsAssalarSuspParcRetir	já era numérica (%)	Iguais
%CDsAssalarSemVarRetir	já era numérica (%)	Iguais
%CDsDispens	já era numérica (%)	Iguais
%AuxilDispens	já era numérica (%)	Iguais
TrocaForneced	Binária (Não /Sim)	0; 1
TrocaLab	Binária (Não /Sim)	0; 1
TrocaTabela	Binária (Não /Sim)	0; 1
FinancExt	Binária (Não /Sim)	0; 1
OutrasMedidas	Binária (Não /Sim)	0; 1
DesequiFinanc	SustentávelPassag; AlgumasMudanç; GrdsMudanç; Venda; Fechamt	0; 1; 2; 3; 4
PlanejSuspTotSocPropr	Binária (Não /Sim)	0; 1
PlanejSuspParcSocPropr	Binária (Não /Sim)	0; 1
PlanejDesligSocPropr	Binária (Não /Sim)	0; 1
PlanejSuspTotLocatPrest	Binária (Não /Sim)	0; 1
PlanejSuspParcLocatPrest	Binária (Não /Sim)	0; 1
PlanejDesligLocatPrest	Binária (Não /Sim)	0; 1
PlanejSuspTotAssalar	Binária (Não /Sim)	0; 1
PlanejSuspParcAssalar	Binária (Não /Sim)	0; 1
PlanejDesligAssalar	Binária (Não /Sim)	0; 1
EmprestExt	Binária (Não /Sim)	0; 1
FechtoVenda	Binária (Não /Sim)	0; 1
MudModeloNeg	Binária (Não /Sim)	0; 1
ExpectativasFuturas	já era numérica (%)	Iguais
ComentOpc	LongChar (mas que não entrarão nas análises, ou apenas servirão como exemplo)	Iguais

APÊNDICE C – O workflow completo da pesquisa

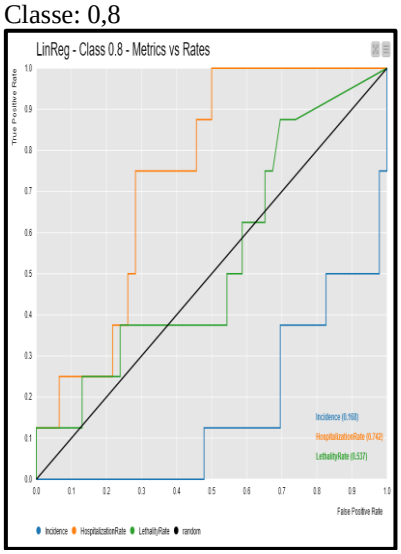


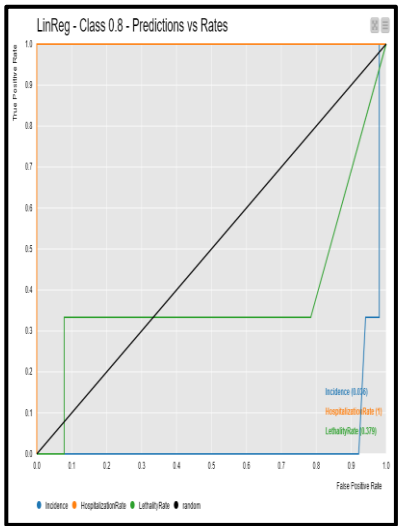
APÊNDICE D – Curvas ROC e AUCs para os diferentes algoritmos

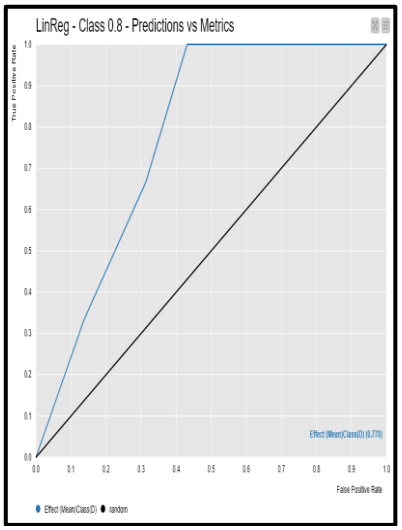
Gráficos 20(a)-(c) – *LinReg*: Curvas ROC (e AUC) para as diferentes classes

(a) Métricas *vs* Taxas **(b) Predições *vs* Taxas** **(c) Predições *vs* Métricas**

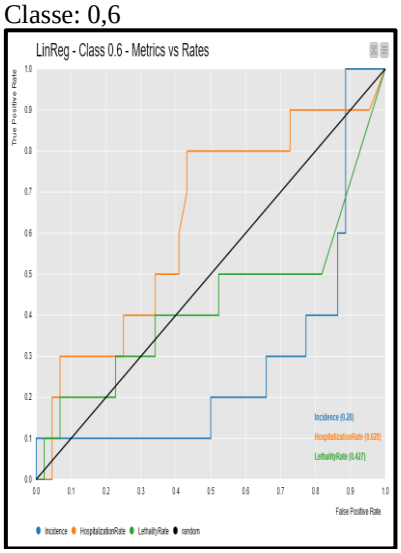
Classe: 0,8

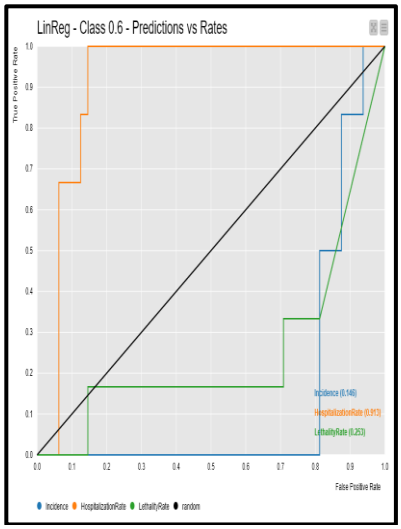


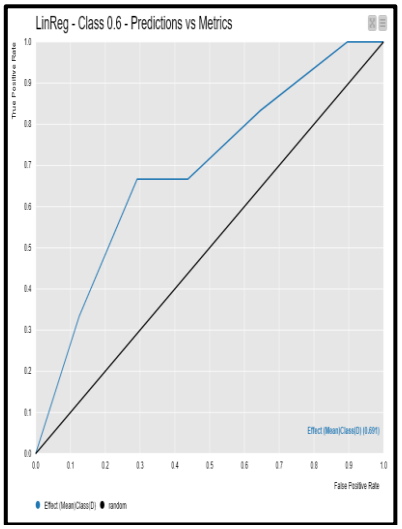




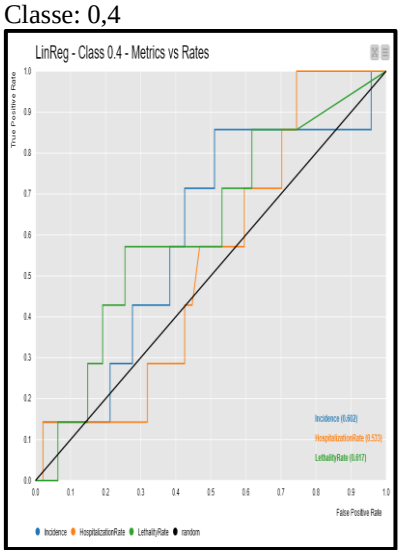
Classe: 0,6

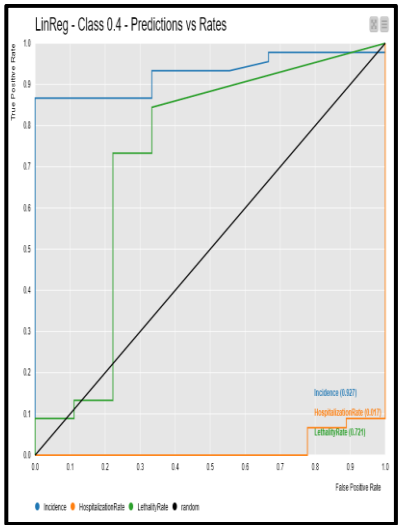


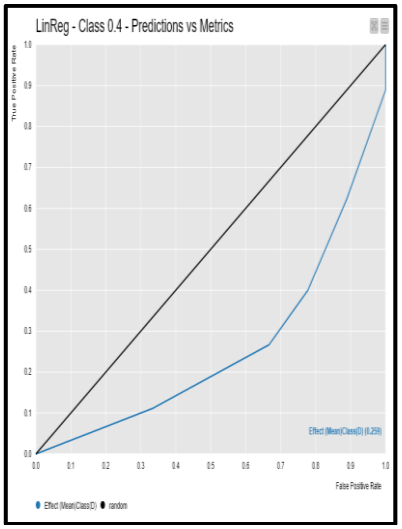




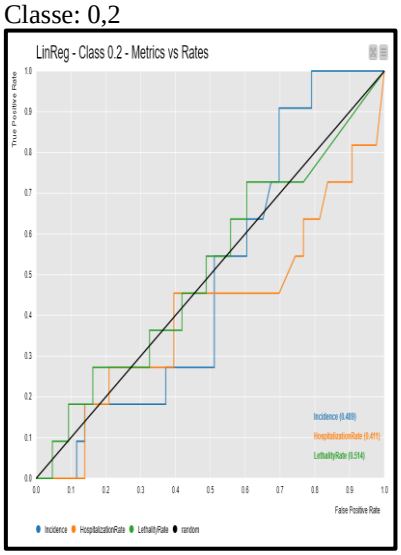
Classe: 0,4





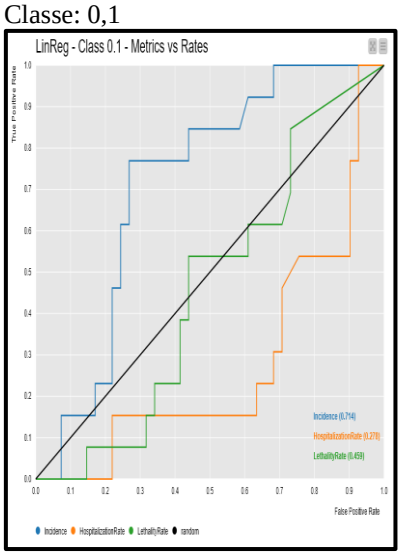


Classe: 0,2



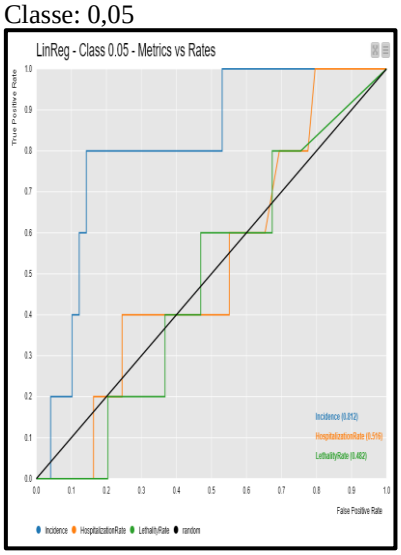
SEM PREDIÇÕES SEM PREDIÇÕES

Classe: 0,1



SEM PREDIÇÕES SEM PREDIÇÕES

Classe: 0,05

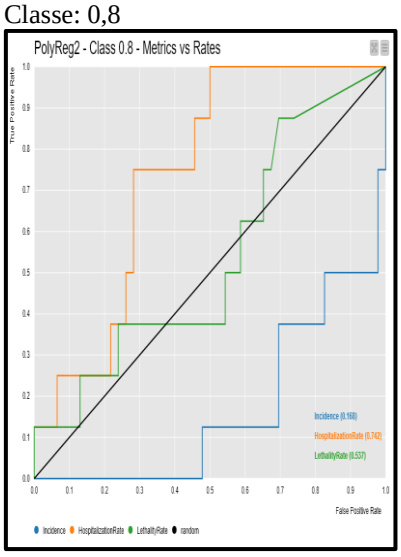


SEM PREDIÇÕES SEM PREDIÇÕES

Gráficos 21(a)-(c) – *PolyReg*²: Curvas ROC (e AUC) para as diferentes classes

(a) Métricas *vs* Taxas **(b) Predições *vs* Taxas** **(c) Predições *vs* Métricas**

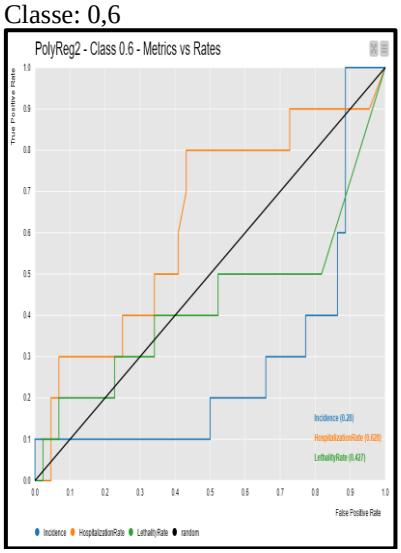
Classe: 0,8

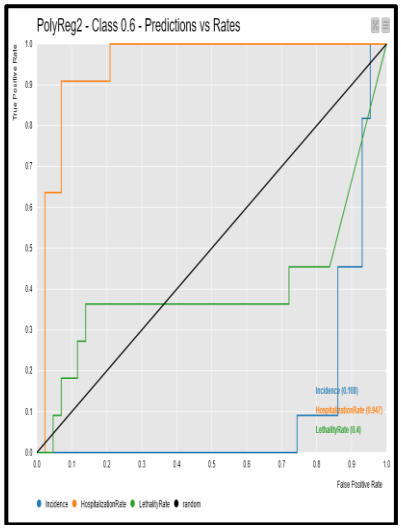


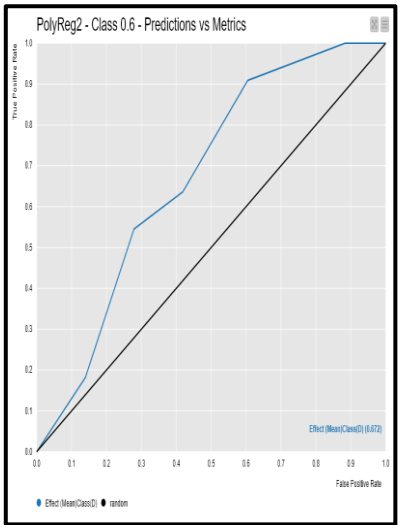
SEM PREDIÇÕES

SEM PREDIÇÕES

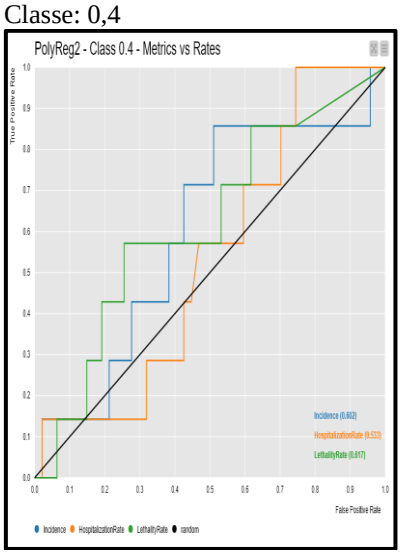
Classe: 0,6

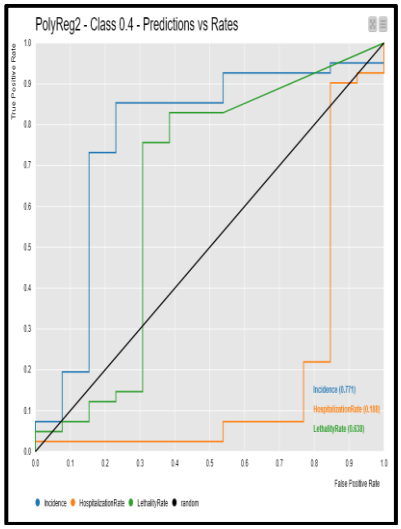


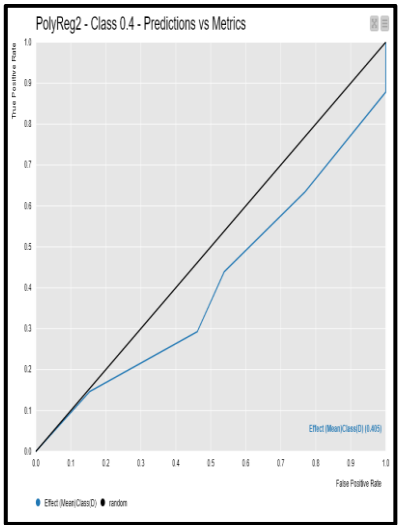




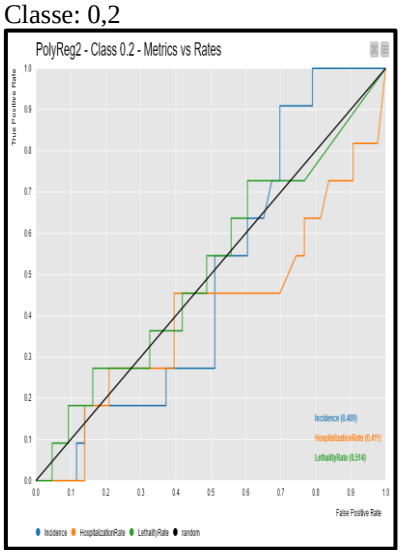
Classe: 0,4

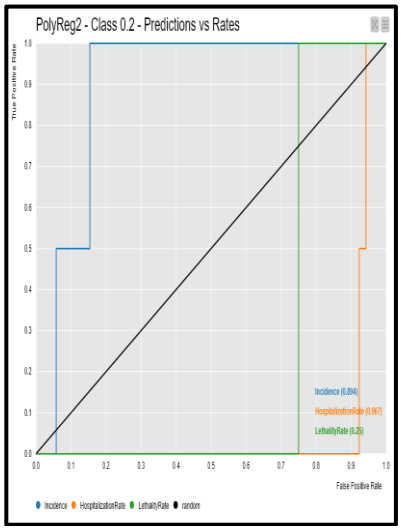


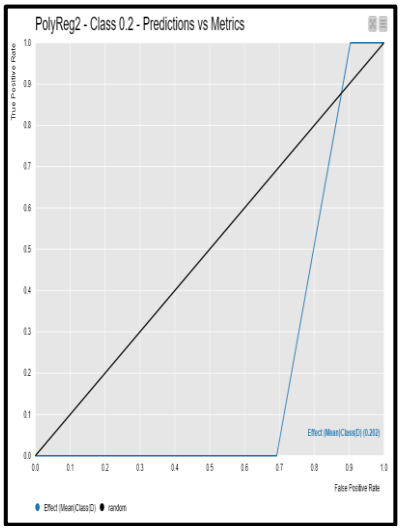




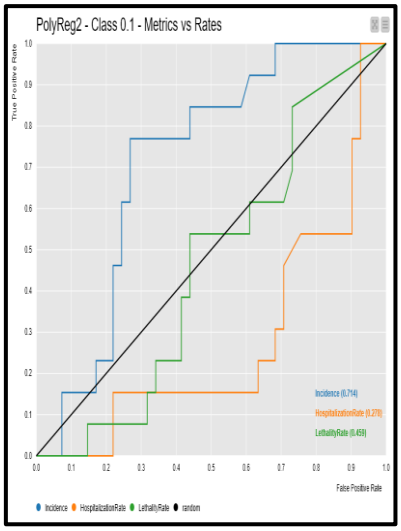
Classe: 0,2







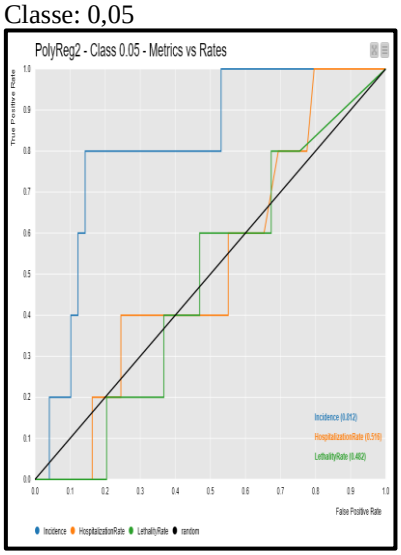
Classe: 0.1



SEM PREDIÇÕES

SEM PREDIÇÕES

Classe: 0,05

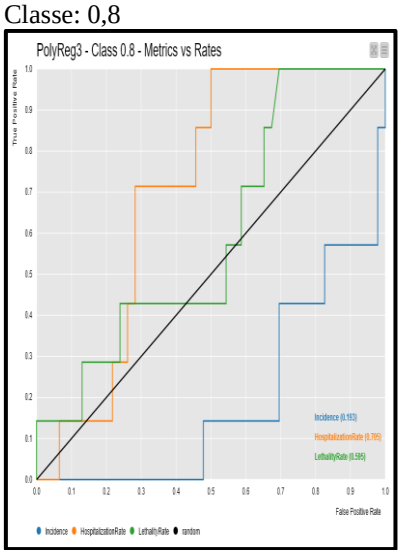


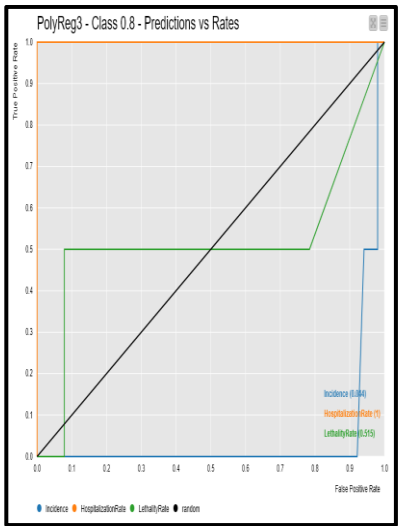
SEM PREDIÇÕES

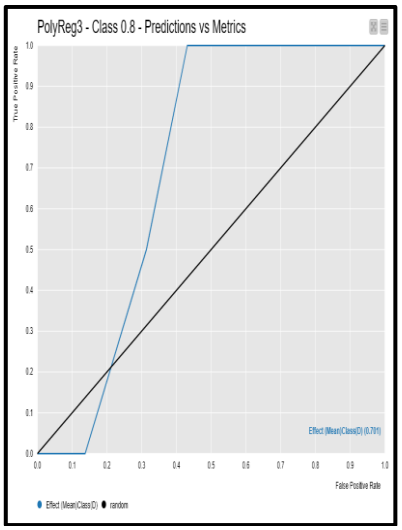
SEM PREDIÇÕES

Gráficos 22(a)-(c) – *PolyReg*³: Curvas ROC (e AUC) para as diferentes classes
(a) Métricas *vs* Taxas (b) Predições *vs* Taxas (c) Predições *vs* Métricas

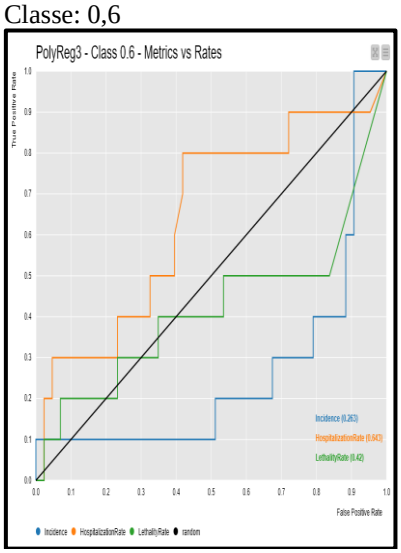
Classe: 0,8

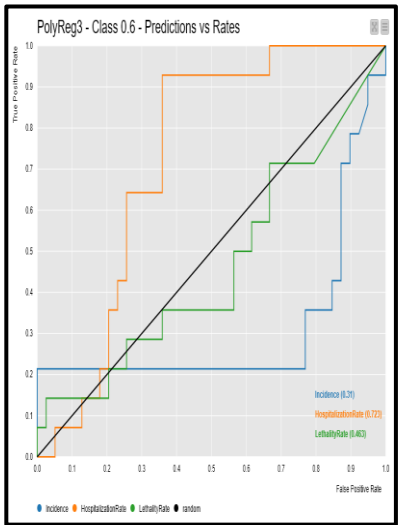


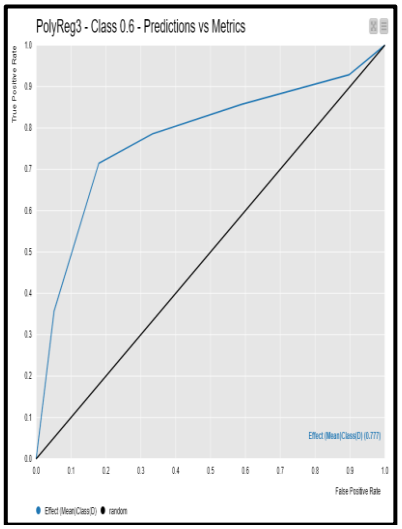




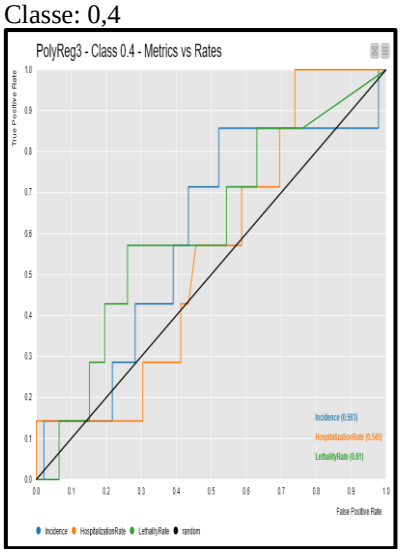
Classe: 0,6

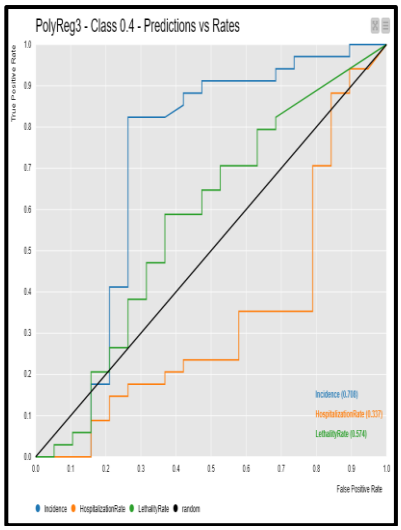


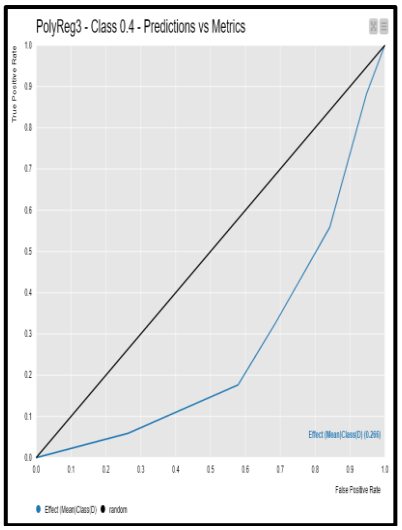




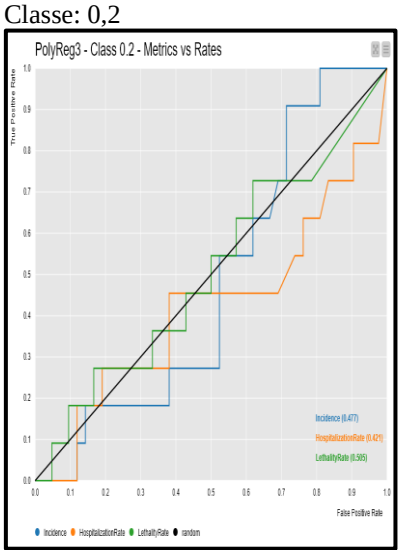
Classe: 0,4

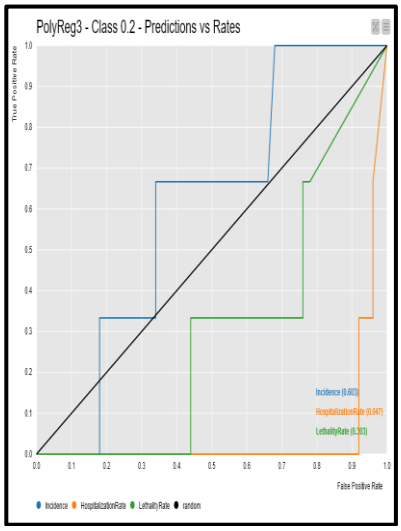


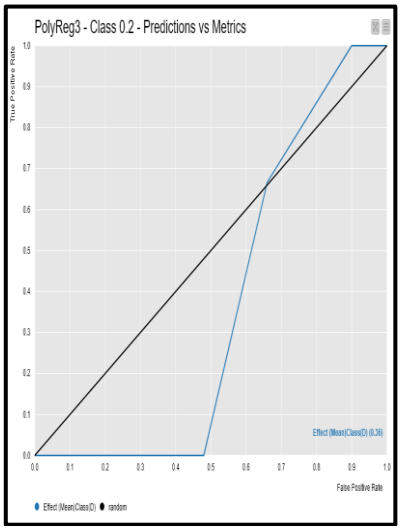




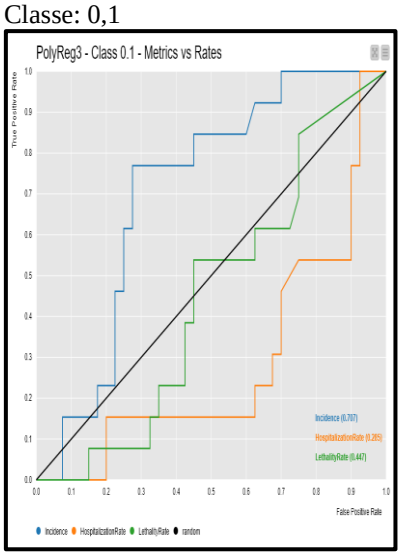
Classe: 0,2







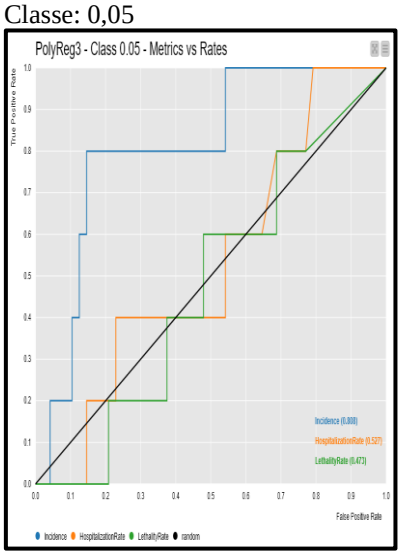
Classe: 0,1



SEM PREDIÇÕES

SEM PREDIÇÕES

Classe: 0,05

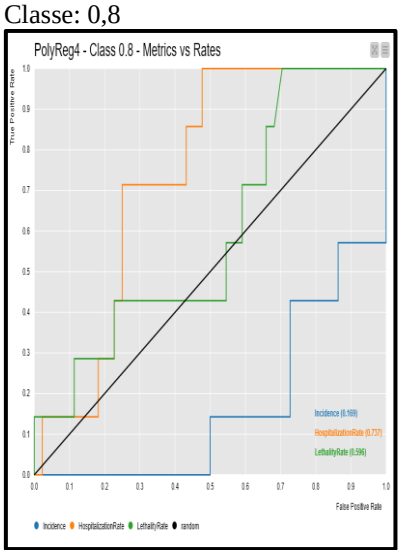


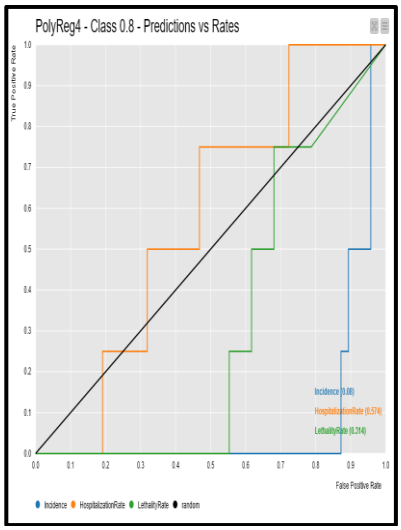
SEM PREDIÇÕES

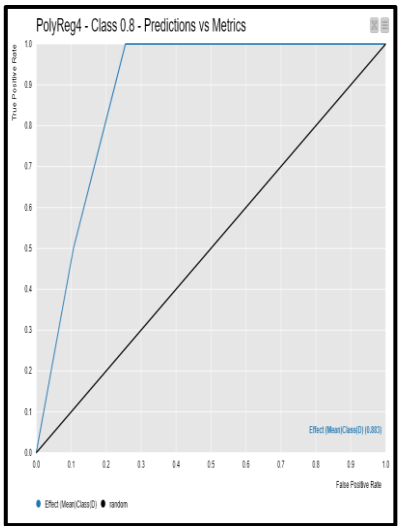
SEM PREDIÇÕES

Gráficos 23(a)-(c) – *PolyReg*^{4}: Curvas ROC (e AUC) para as diferentes classes
(a) Métricas *vs* Taxas (b) Predições *vs* Taxas (c) Predições *vs* Métricas

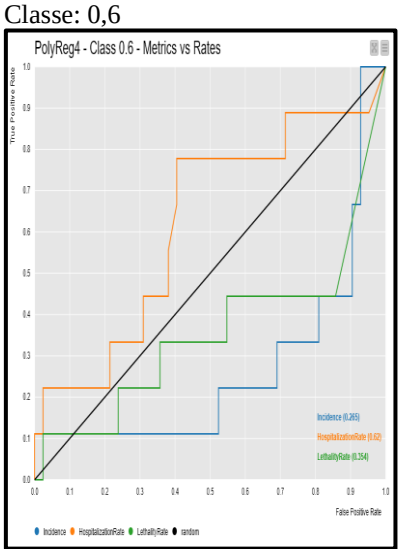
Classe: 0,8

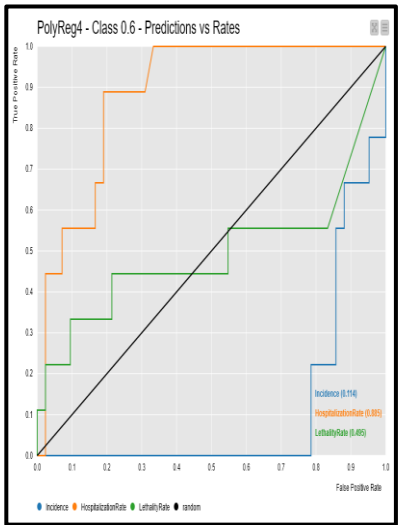


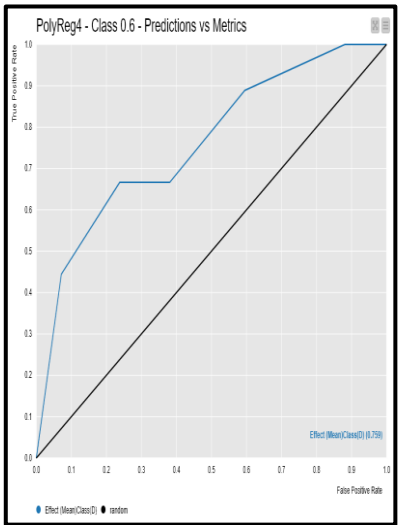




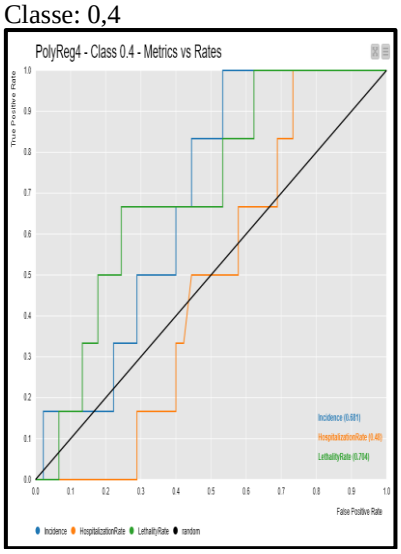
Classe: 0,6

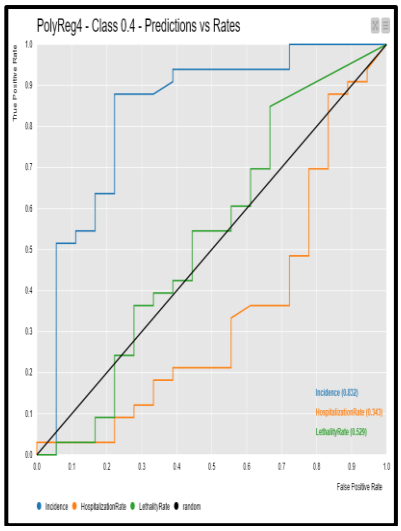


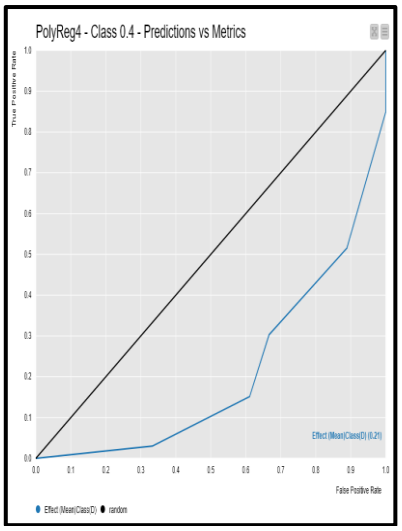




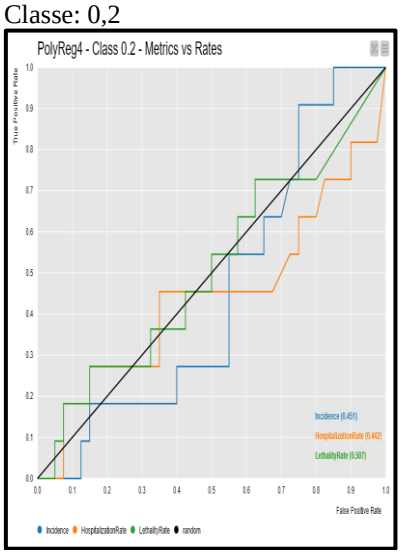
Classe: 0,4

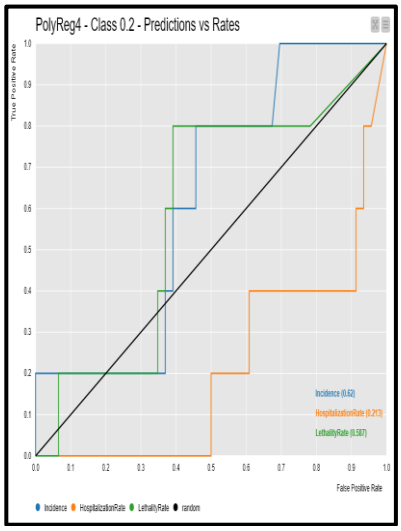


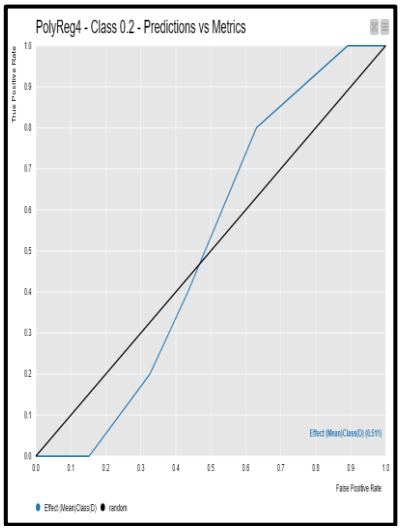




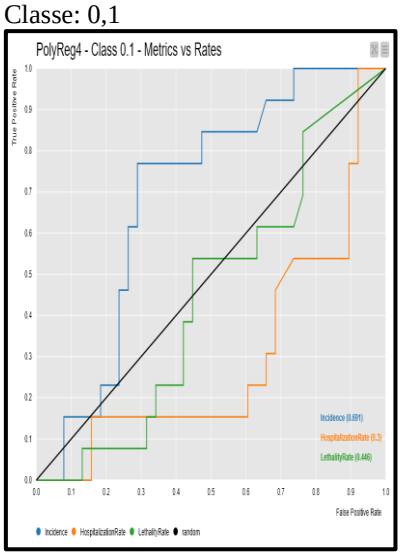
Classe: 0,2







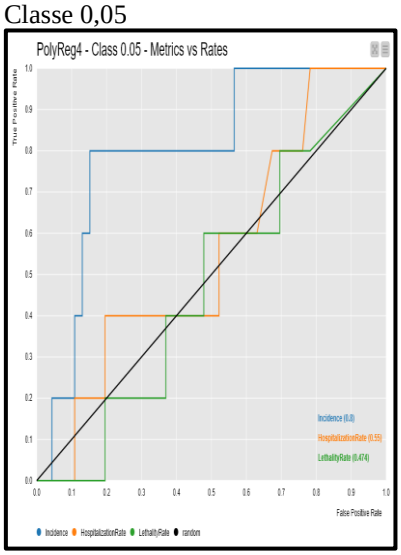
Classe: 0,1



SEM PREDIÇÕES

SEM PREDIÇÕES

Classe 0,05



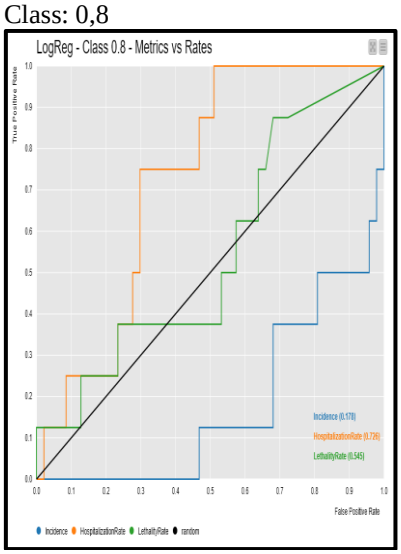
SEM PREDIÇÕES

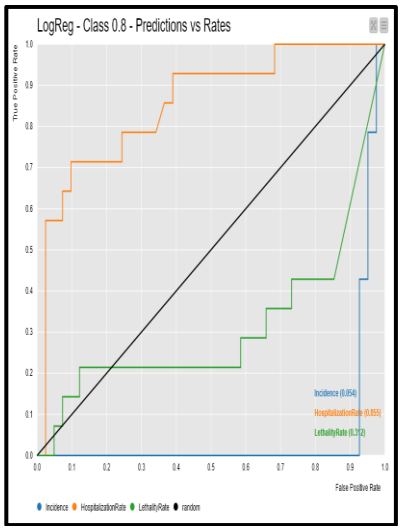
SEM PREDIÇÕES

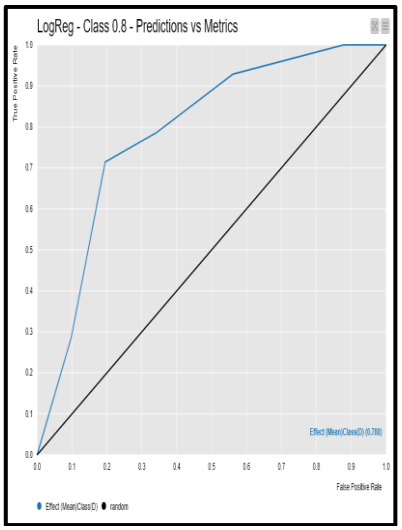
Gráficos 24(a)-(c) – *LogReg*: Curvas ROC (e AUC) para as diferentes classes

(a) Métricas vs Taxas **(b) Predições vs Taxas** **(c) Predições vs Métricas**

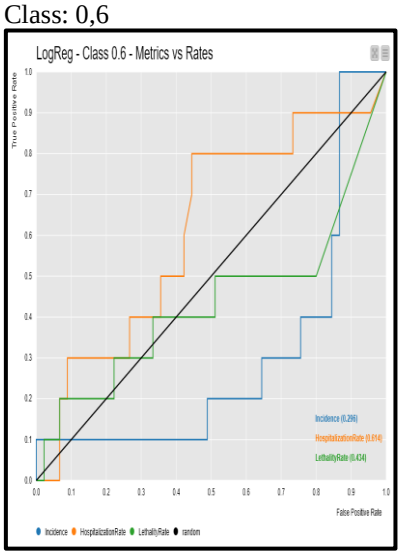
Class: 0,8







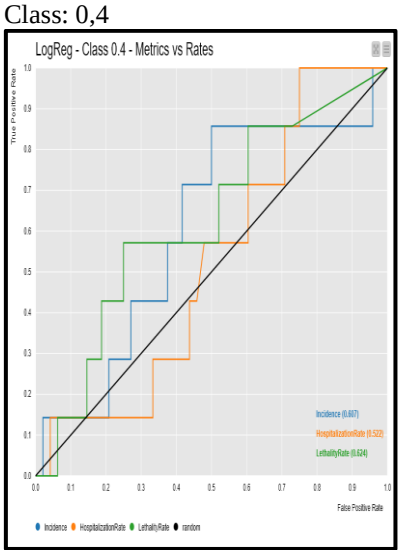
Class: 0,6

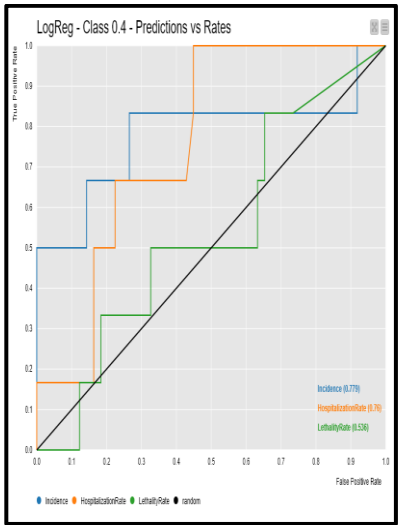


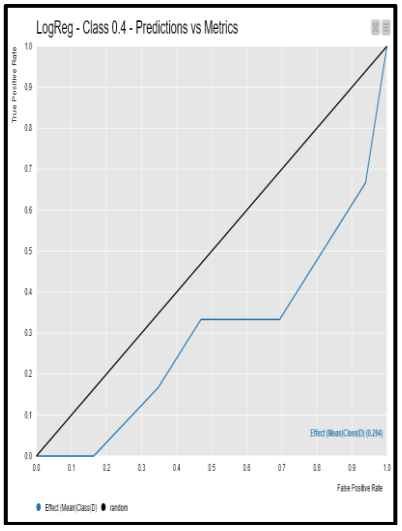
SEM PREDIÇÕES

SEM PREDIÇÕES

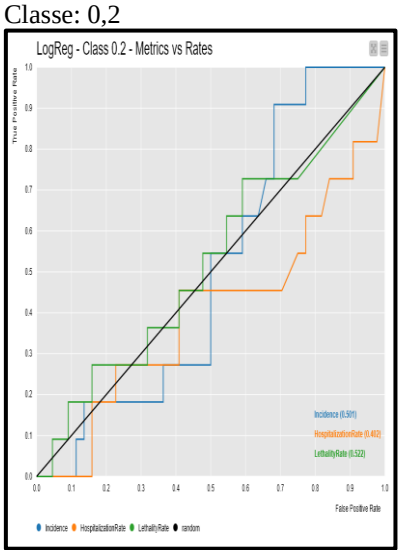
Class: 0,4

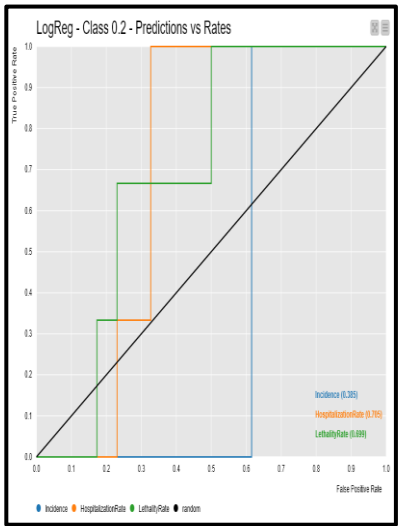


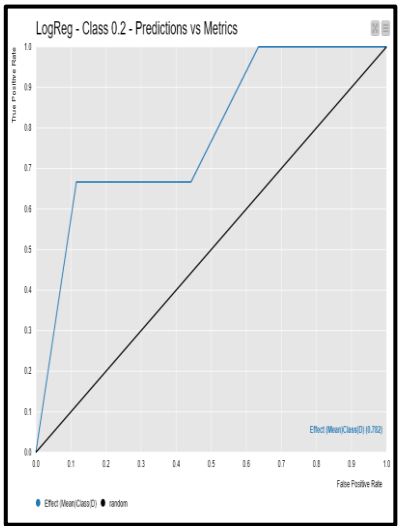




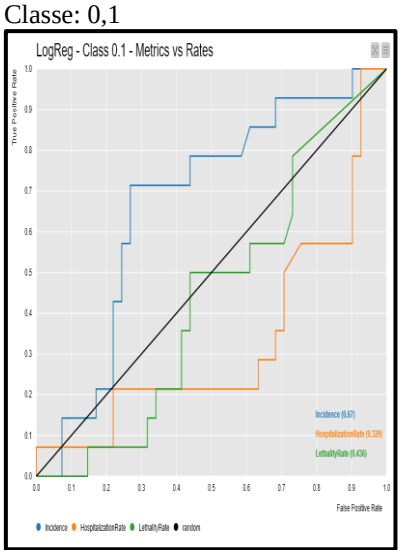
Classe: 0,2

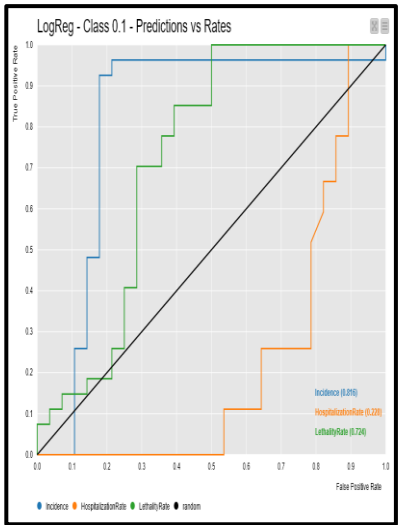


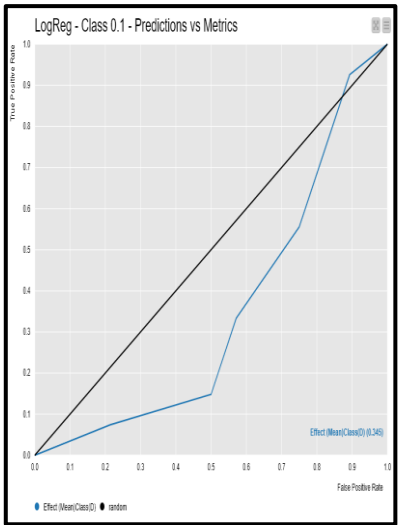




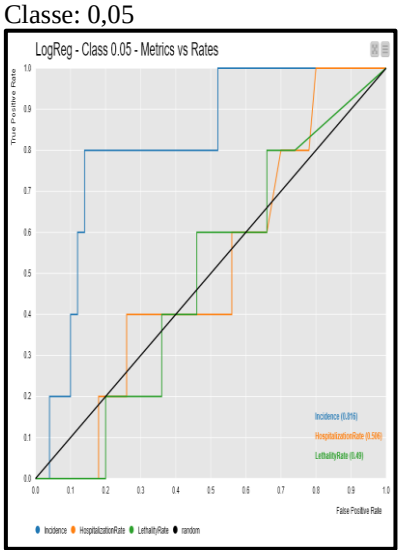
Classe: 0,1

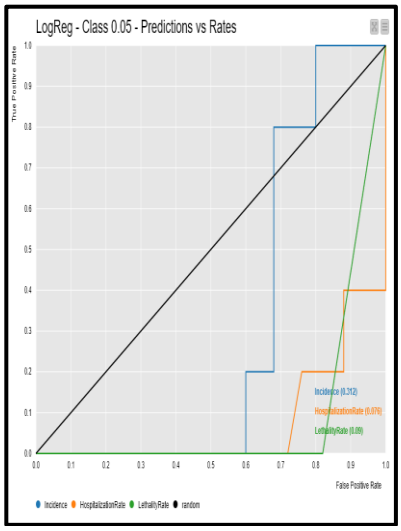


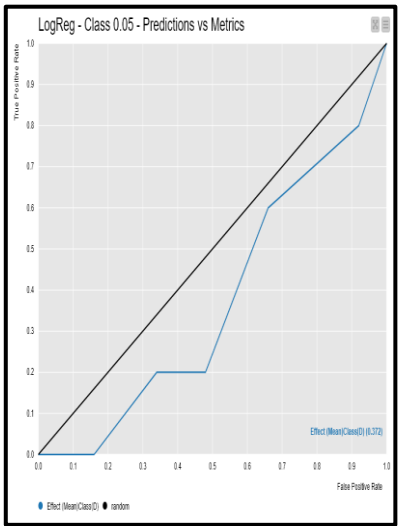




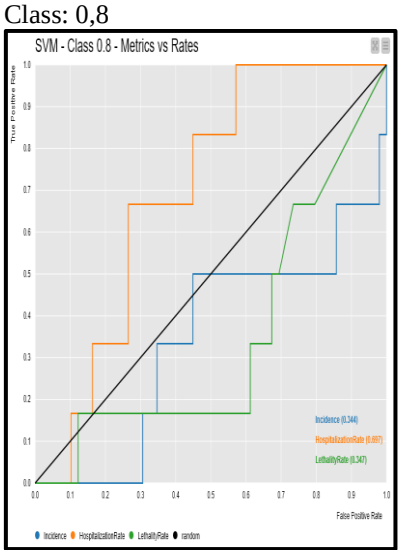
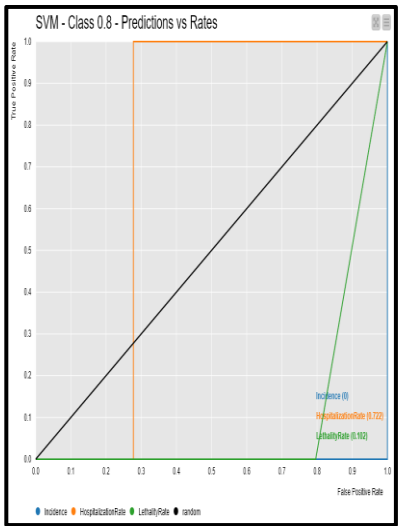
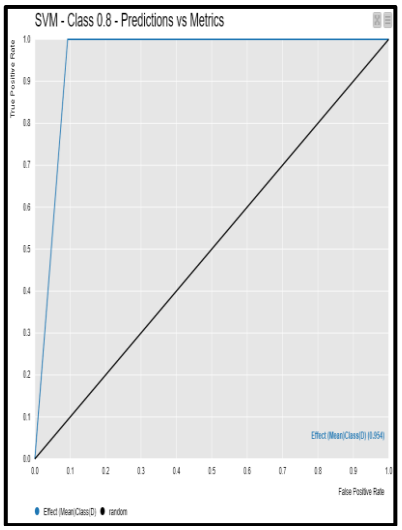
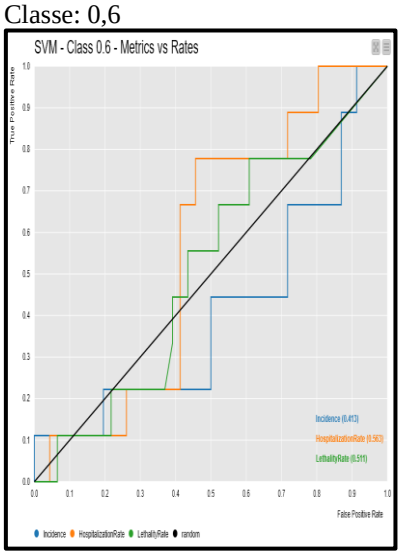
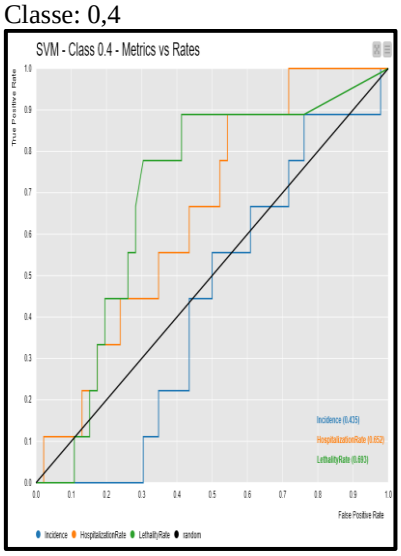
Classe: 0,05



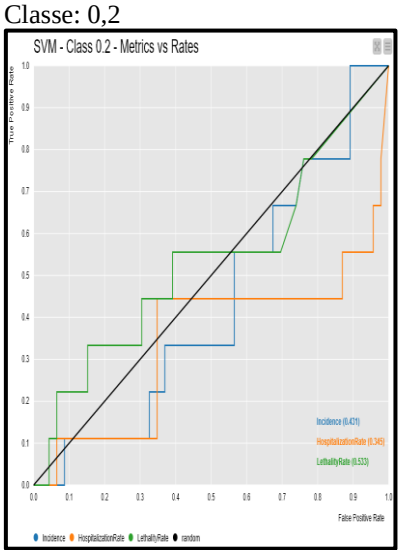


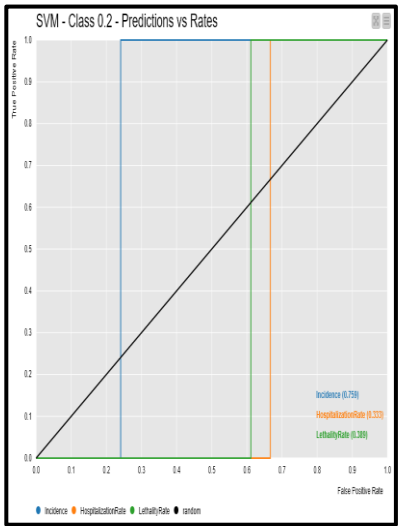


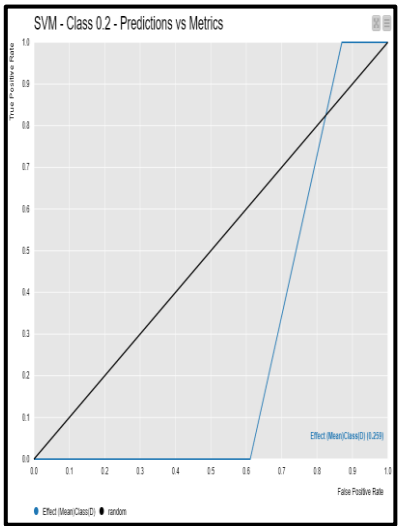
Gráficos 25(a)-(c) – SVM: Curvas ROC (e AUC) para as diferentes classes

(a) Métricas vs Taxas	**(b) Predições vs Taxas**	**(c) Predições vs Métricas**
Class: 0,8		
		
Classe: 0,6		
	SEM PREDIÇÕES	SEM PREDIÇÕES
Classe: 0,4		
	SEM PREDIÇÕES	SEM PREDIÇÕES

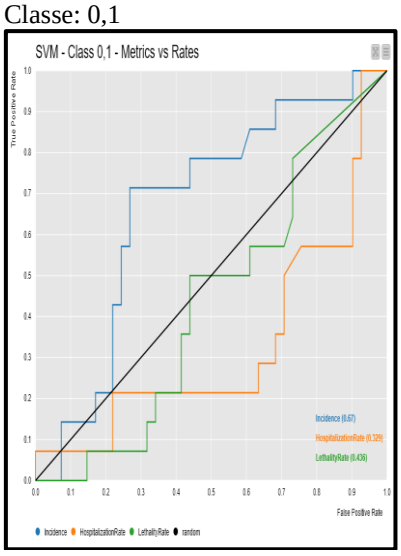
Classe: 0,2

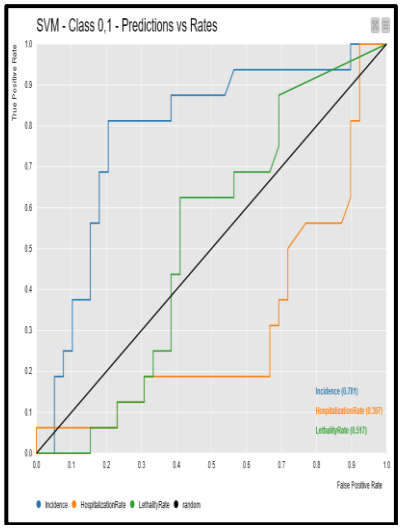


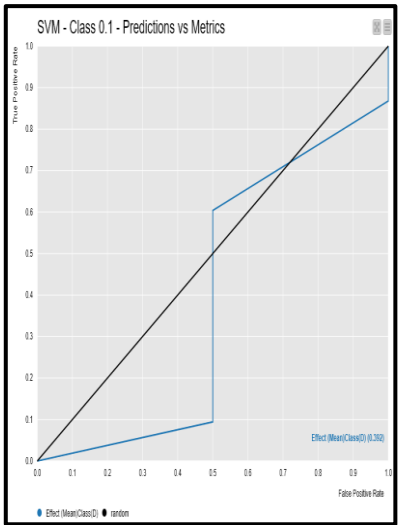




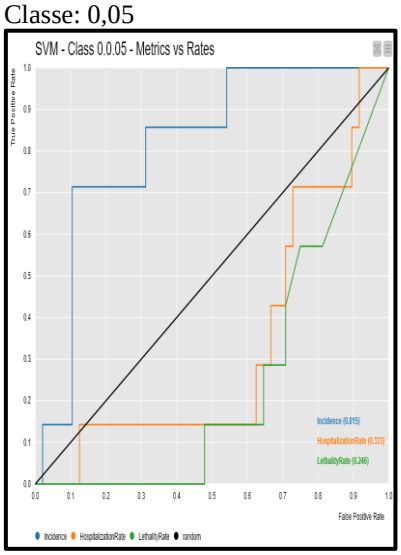
Classe: 0,1







Classe: 0,05



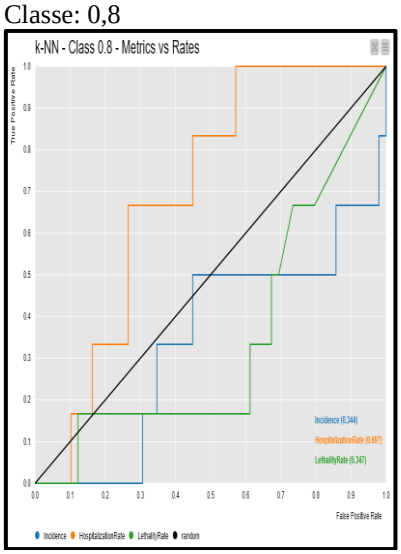
SEM PREDIÇÕES

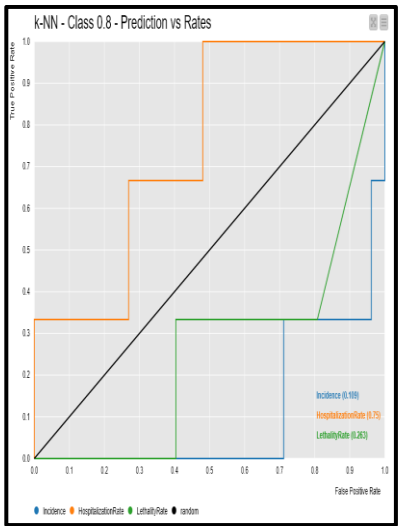
SEM PREDIÇÕES

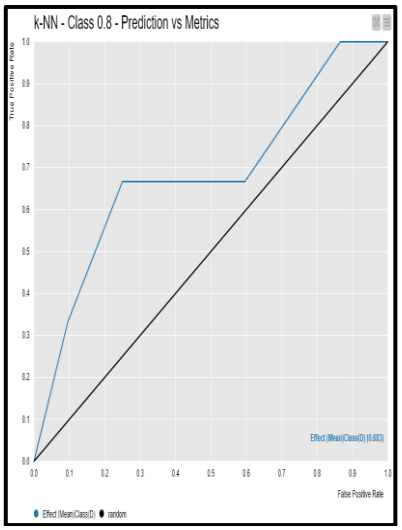
Gráficos 26(a)-(c) – *k*-NN: Curvas ROC (e AUC) para as diferentes classes

(a) Métricas vs Taxas **(b) Predições vs Taxas** **(c) Predições vs Métricas**

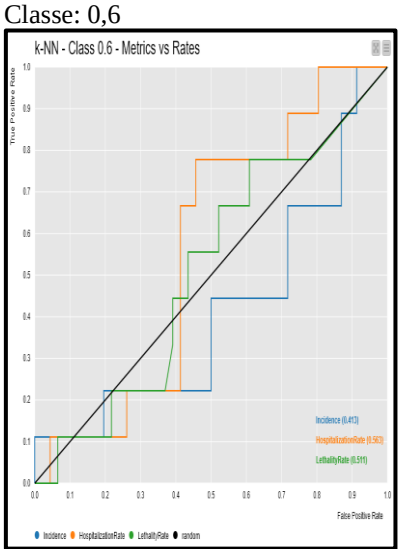
Classe: 0,8

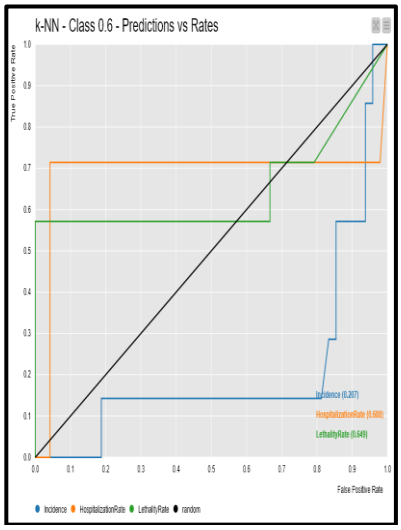


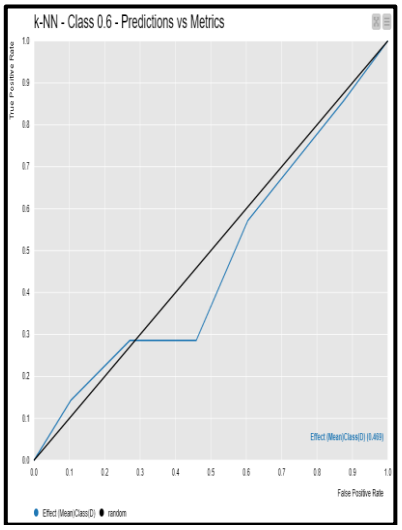




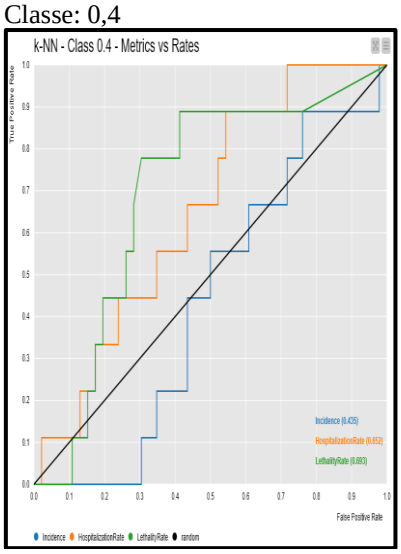
Classe: 0,6

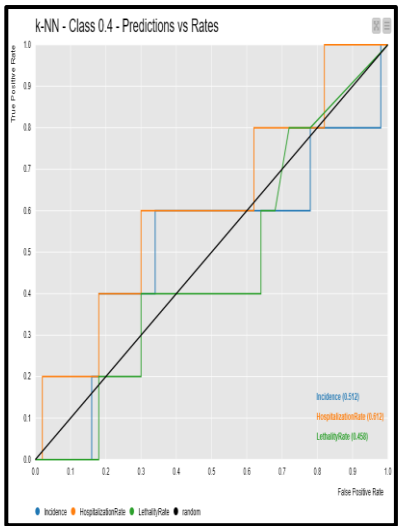


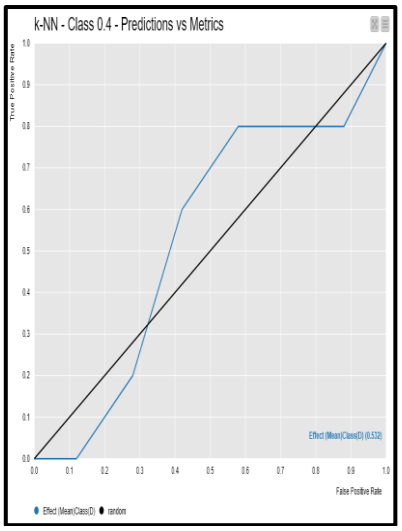




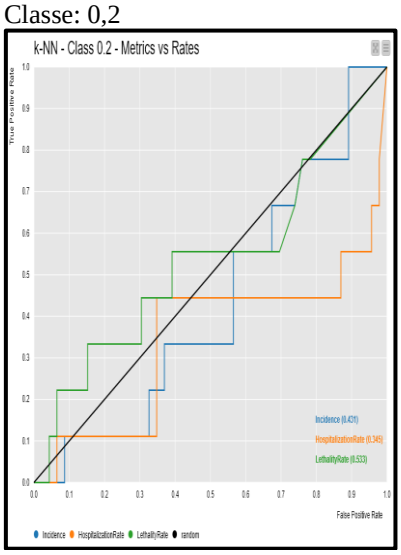
Classe: 0,4

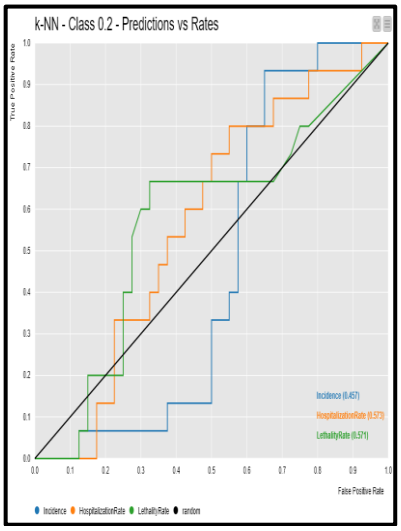


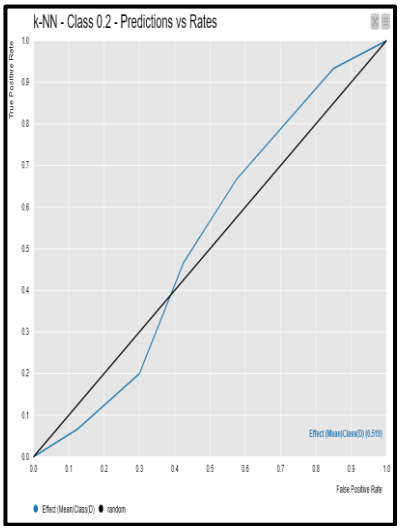




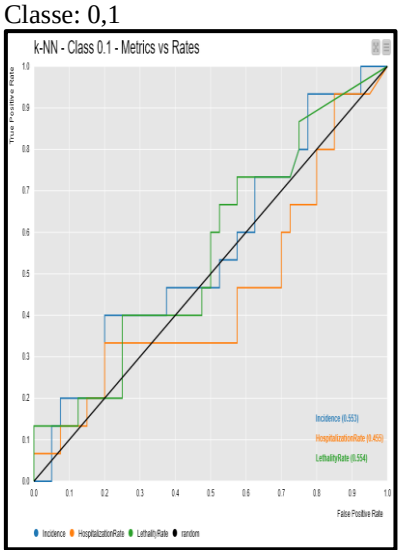
Classe: 0,2

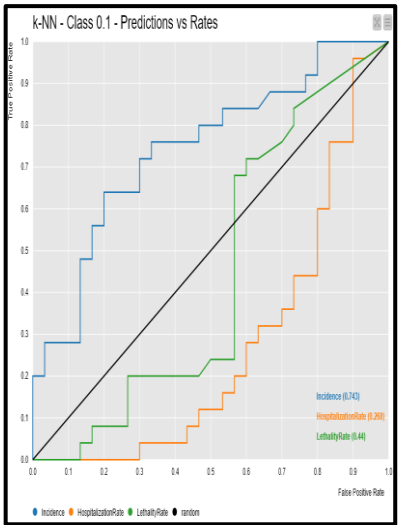


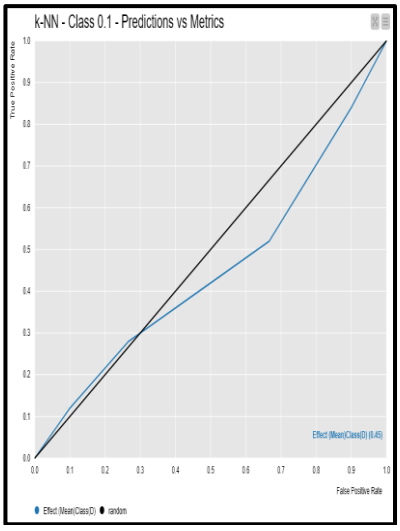




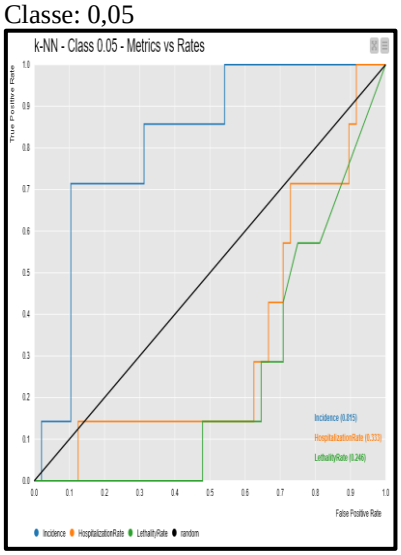
Classe: 0,1







Classe: 0,05



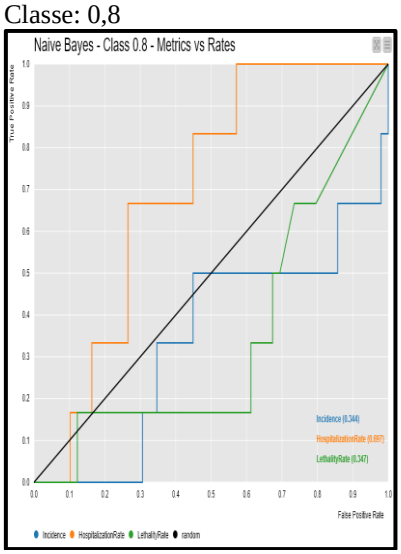
SEM PREDIÇÕES

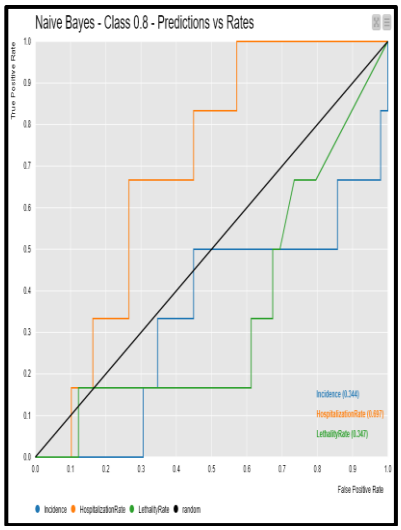
SEM PREDIÇÕES

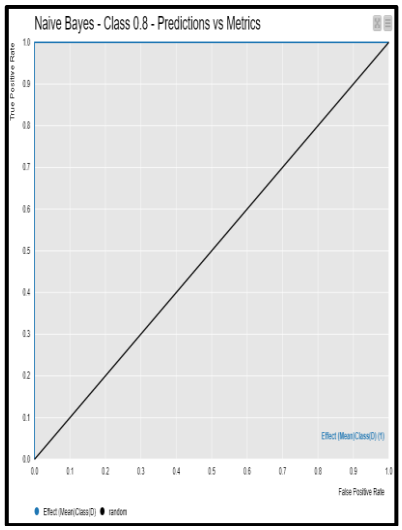
Gráficos 27(a)-(c) – Naïve Bayes: Curvas ROC (e AUC) para as diferentes classes

(a) Métricas vs Taxas **(b) Predições vs Taxas** **(c) Predições vs Métricas**

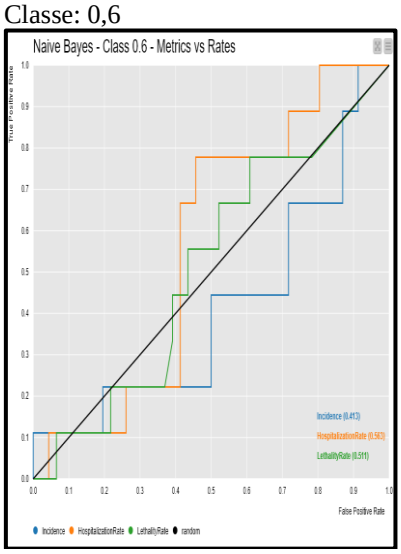
Classe: 0,8

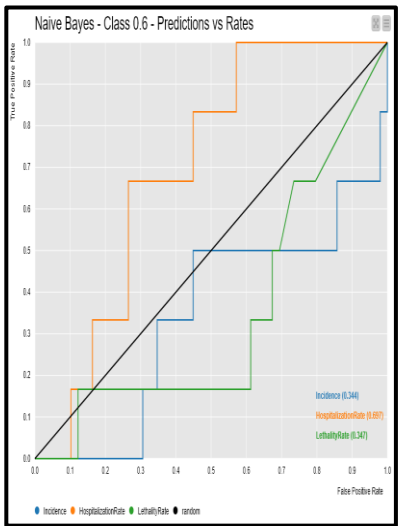


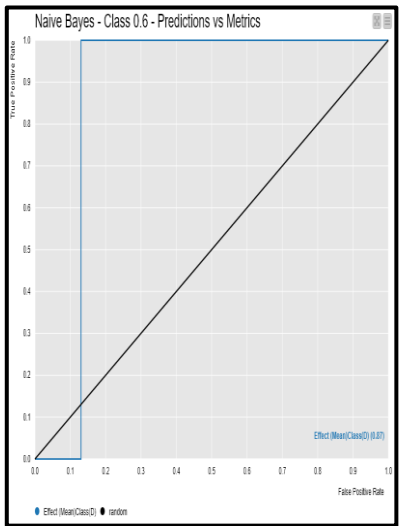




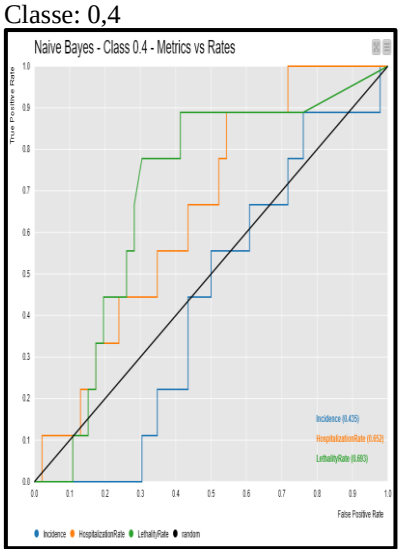
Classe: 0,6

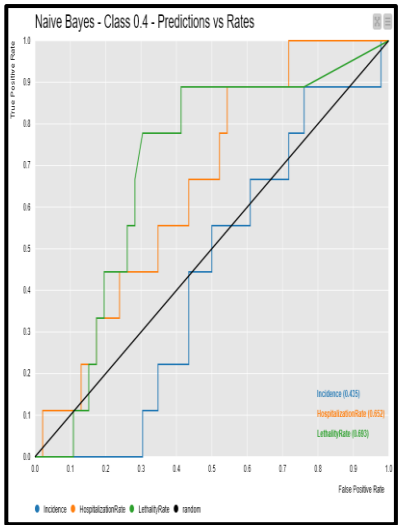


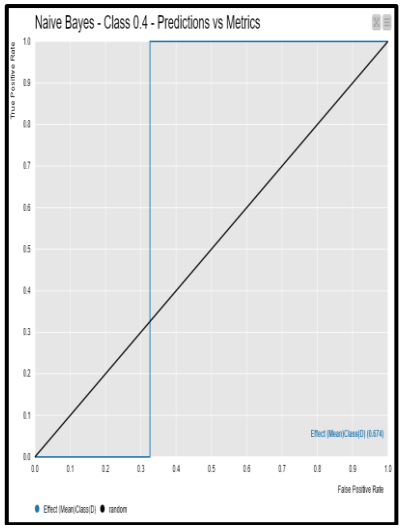




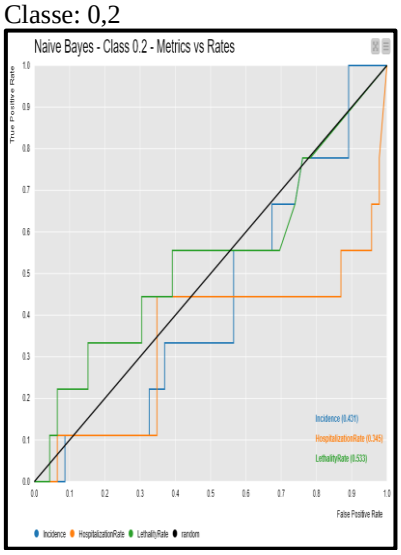
Classe: 0,4

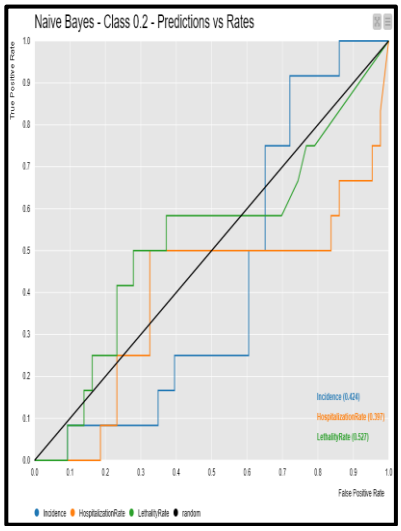


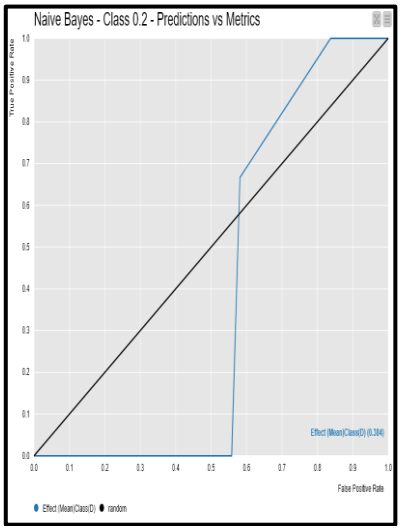




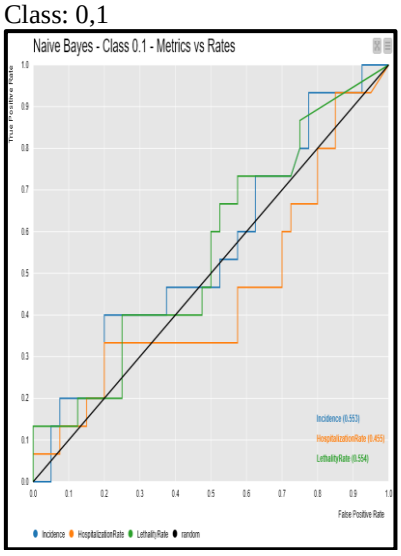
Classe: 0,2

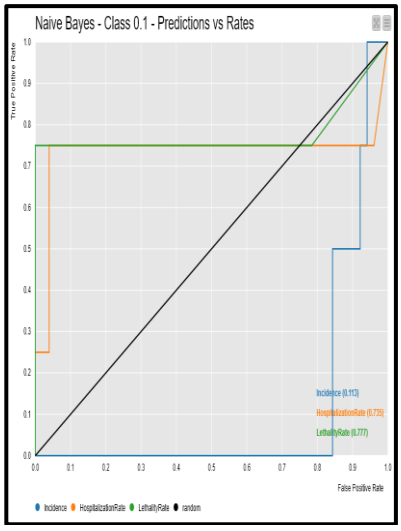


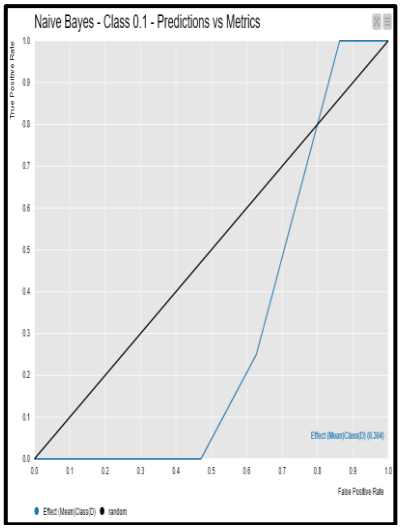




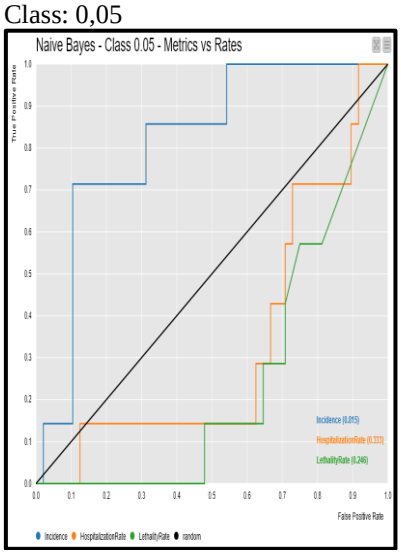
Class: 0,1

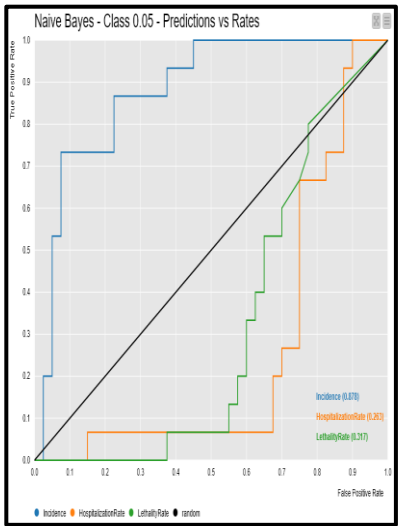


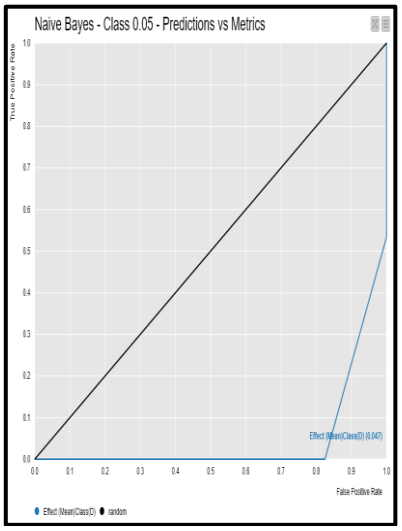




Class: 0,05



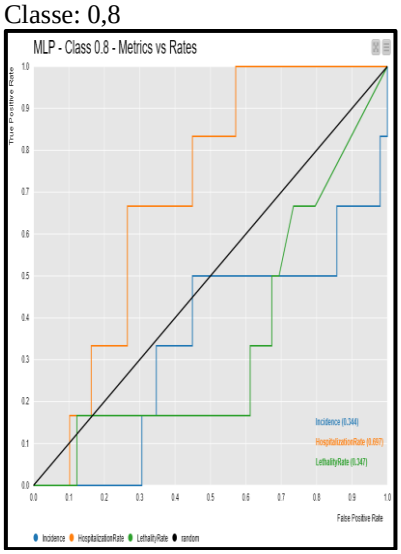


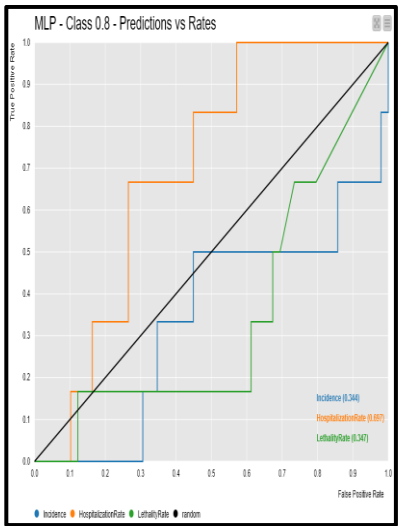


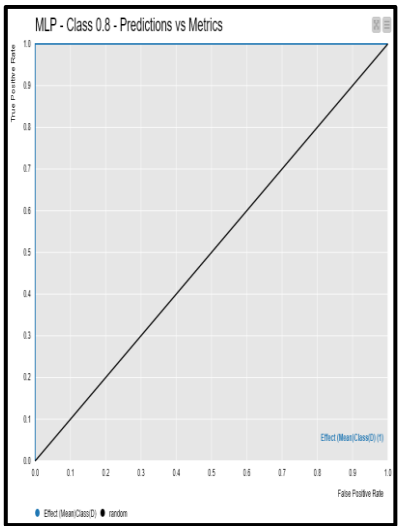
Gráficos 28(a)-(c) – MLP: Curvas ROC (e AUC) para as diferentes classes

(a) Métricas vs Taxas **(b) Predições vs Taxas** **(c) Predições vs Métricas**

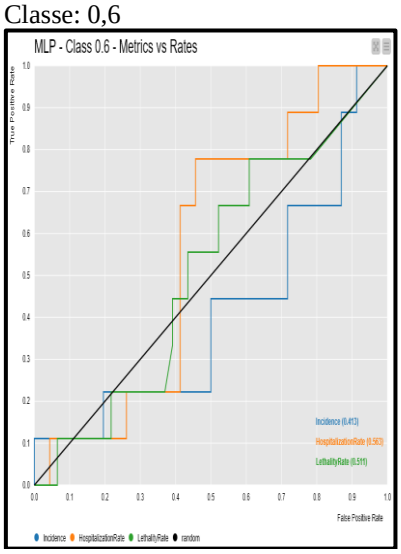
Classe: 0,8

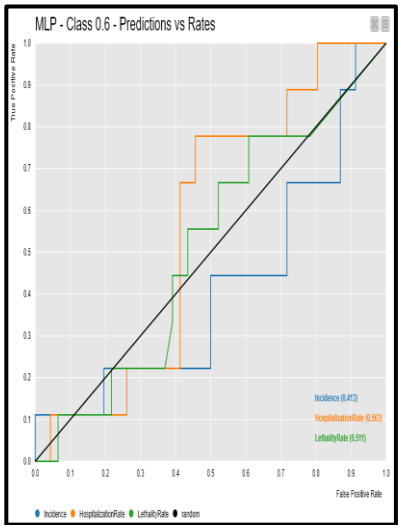


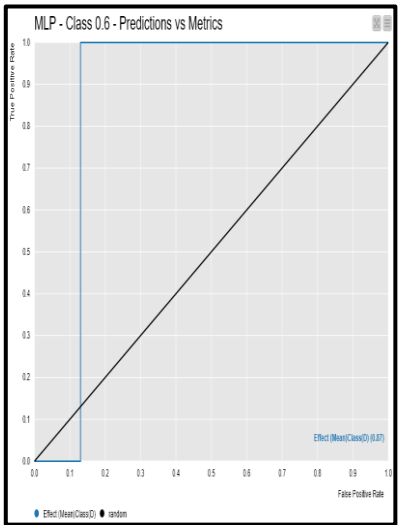




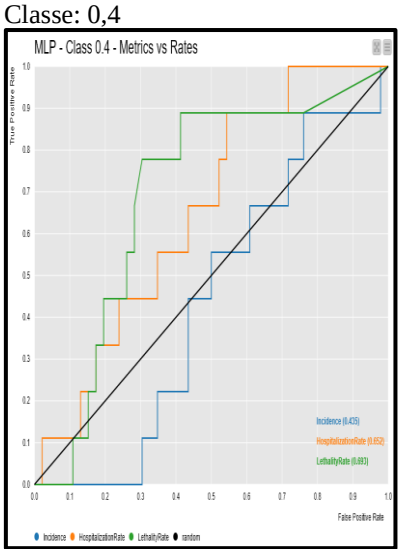
Classe: 0,6

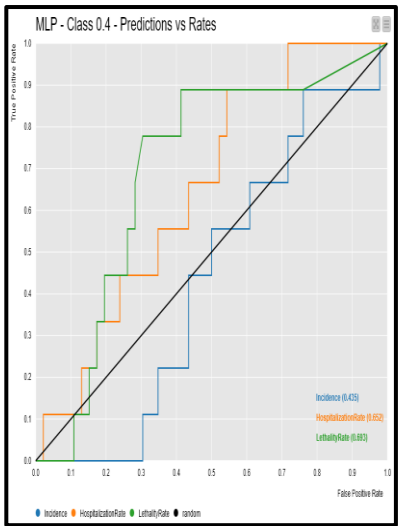


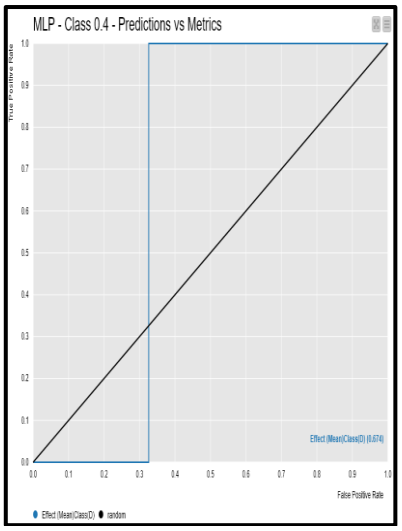




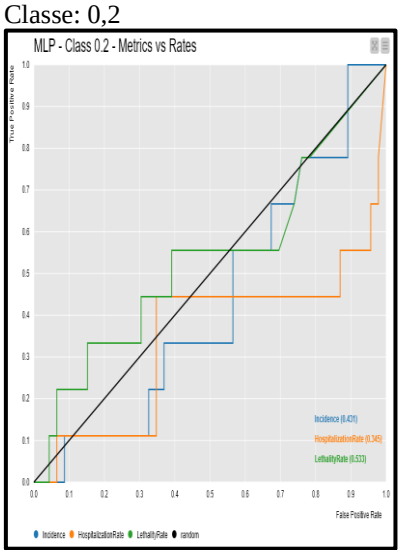
Classe: 0,4

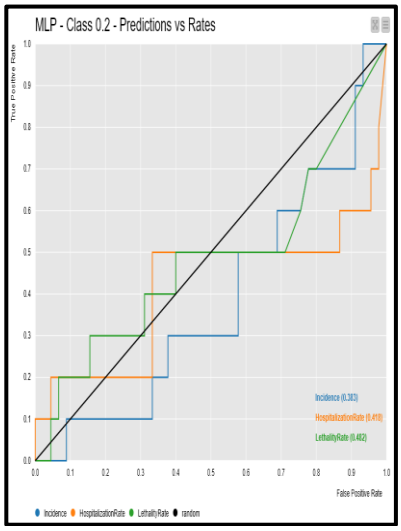


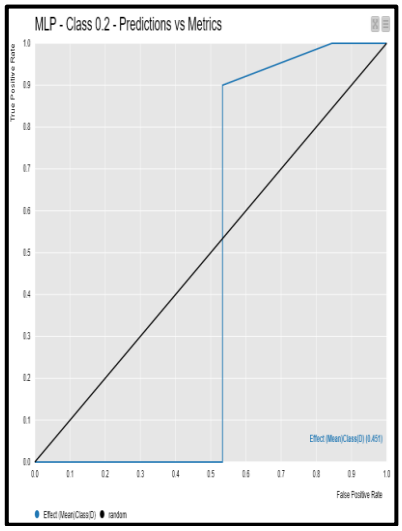




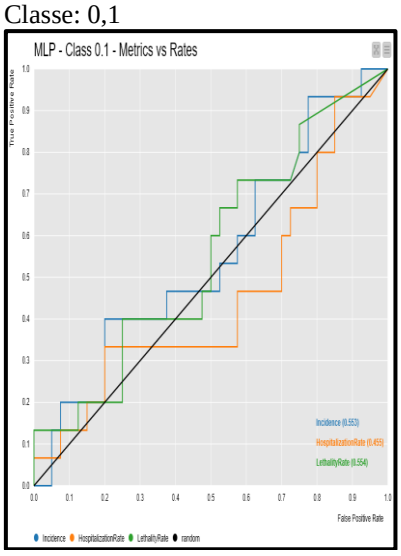
Classe: 0,2

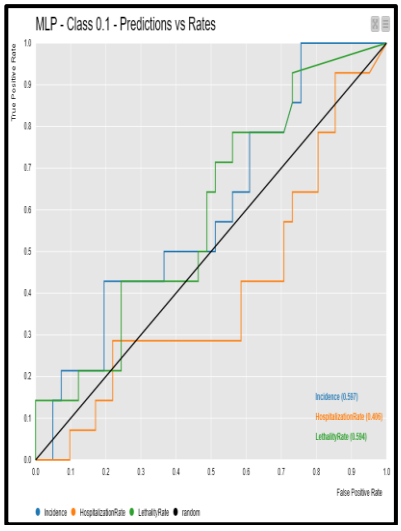


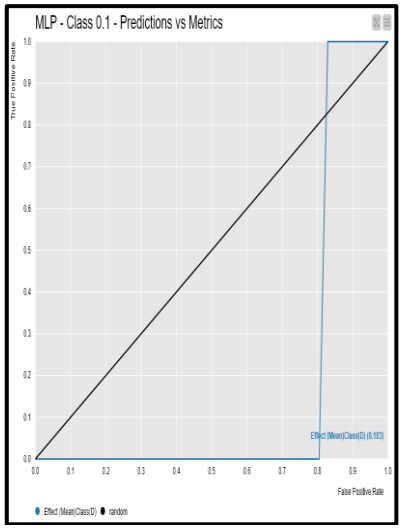




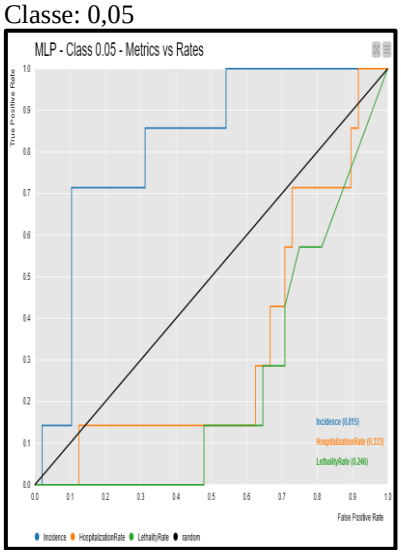
Classe: 0,1

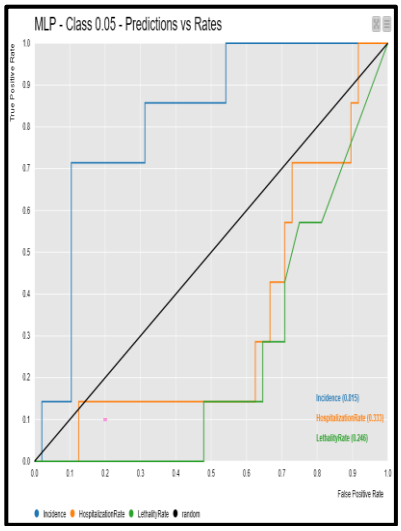


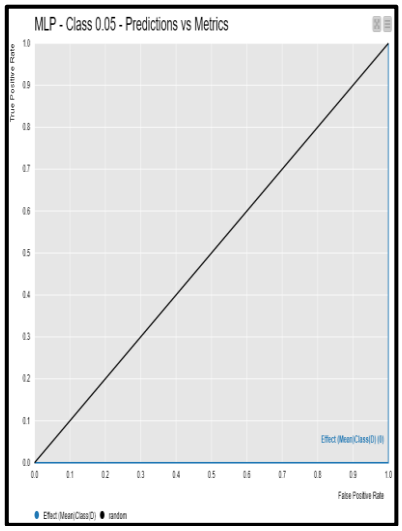




Classe: 0,05



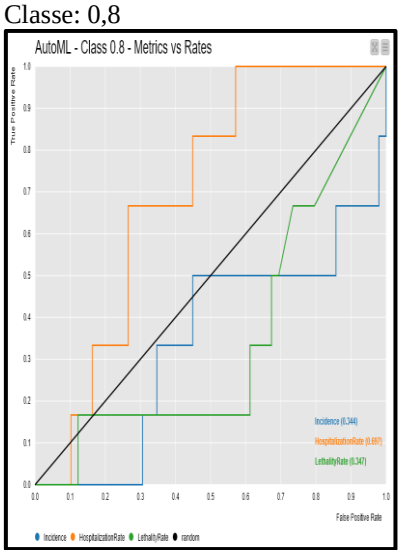


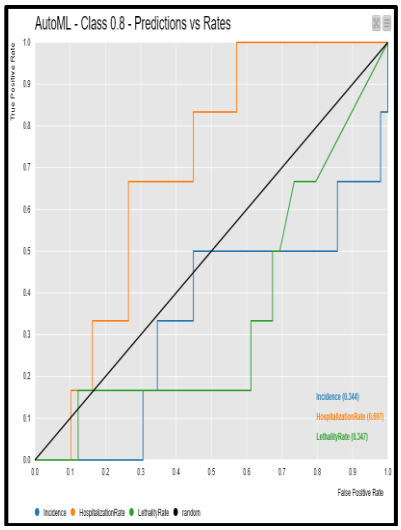


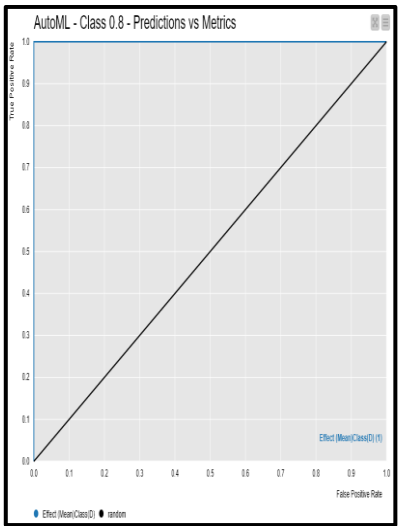
Gráficos 29(a)-(c) – AutoML: Curvas ROC (e AUC) para as diferentes classes

(a) Métricas vs Taxas **(b) Predições vs Taxas** **(c) Predições vs Métricas**

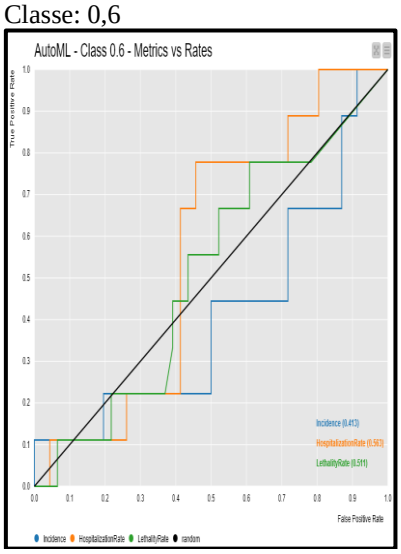
Classe: 0,8

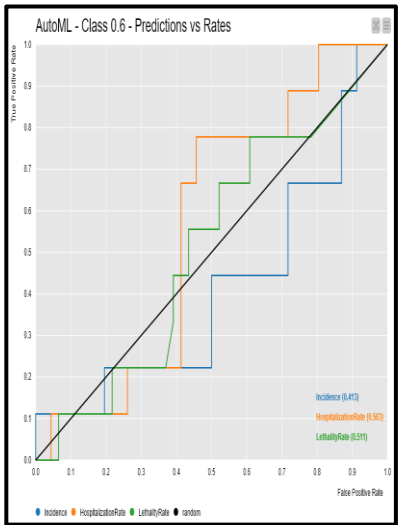


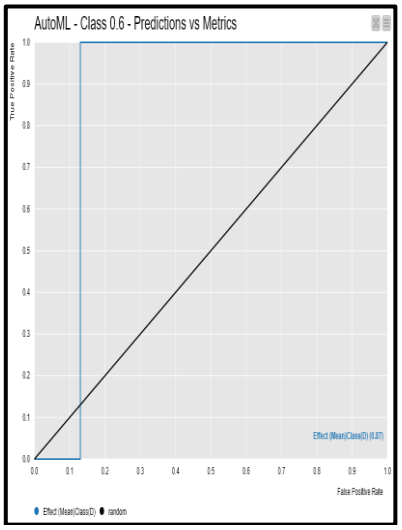




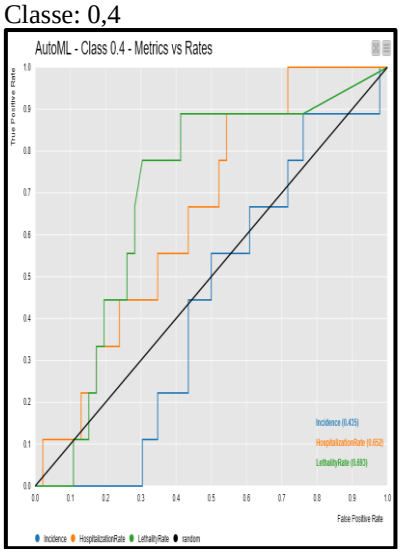
Classe: 0,6

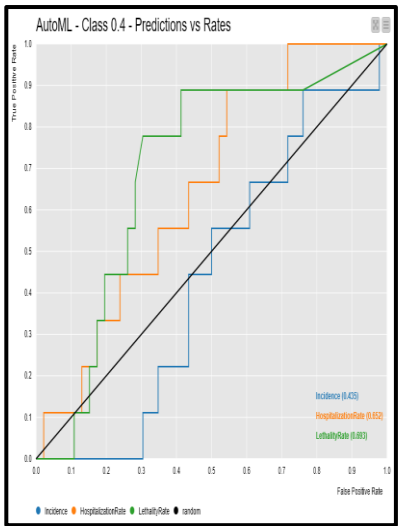


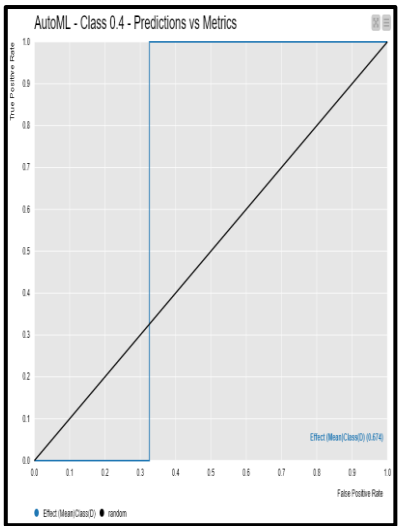




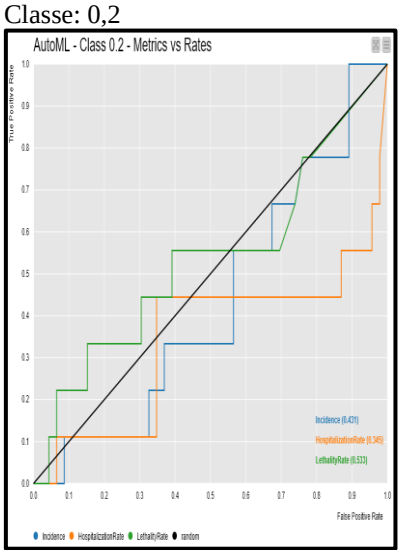
Classe: 0,4

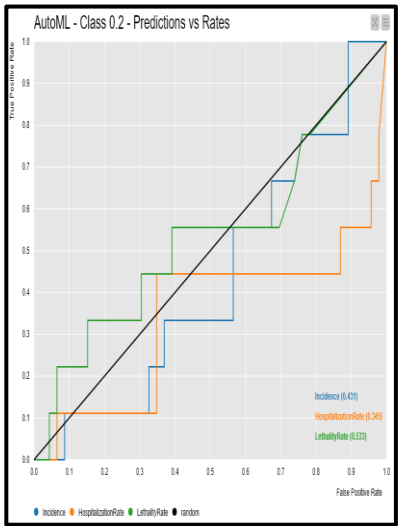


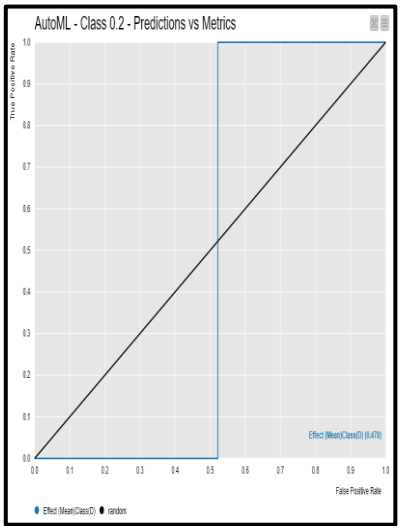




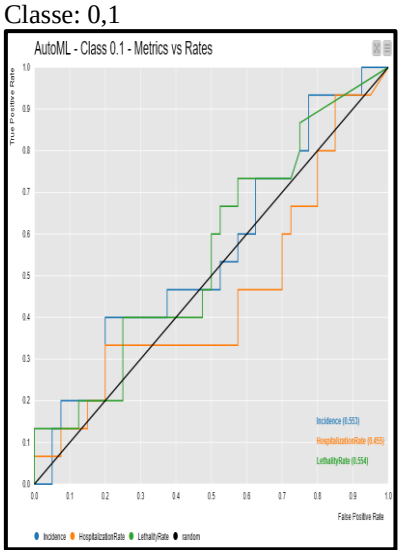
Classe: 0,2

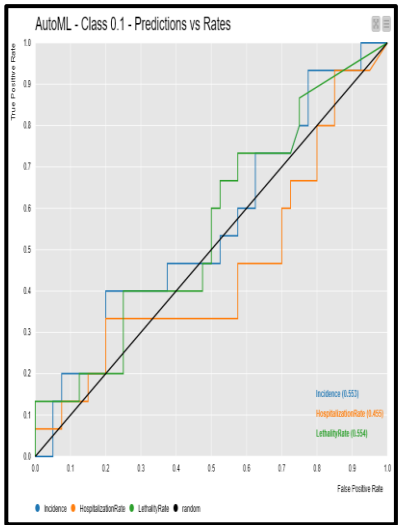


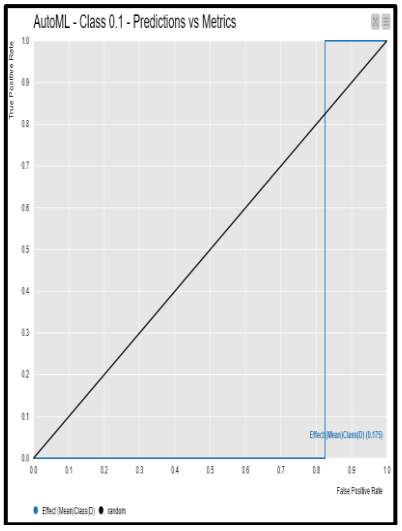




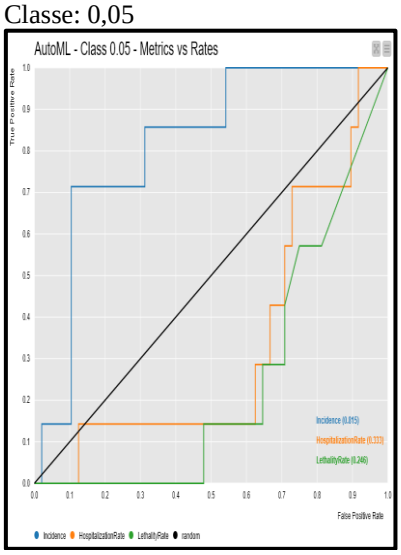
Classe: 0,1

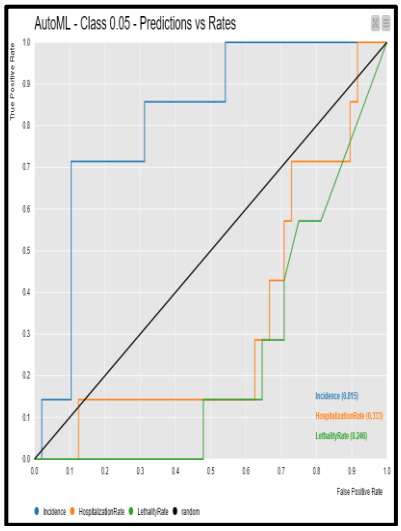


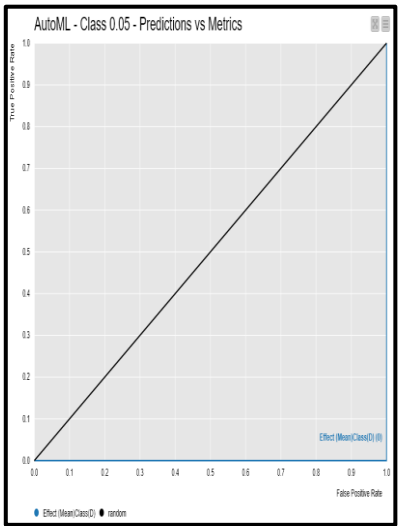




Classe: 0,05

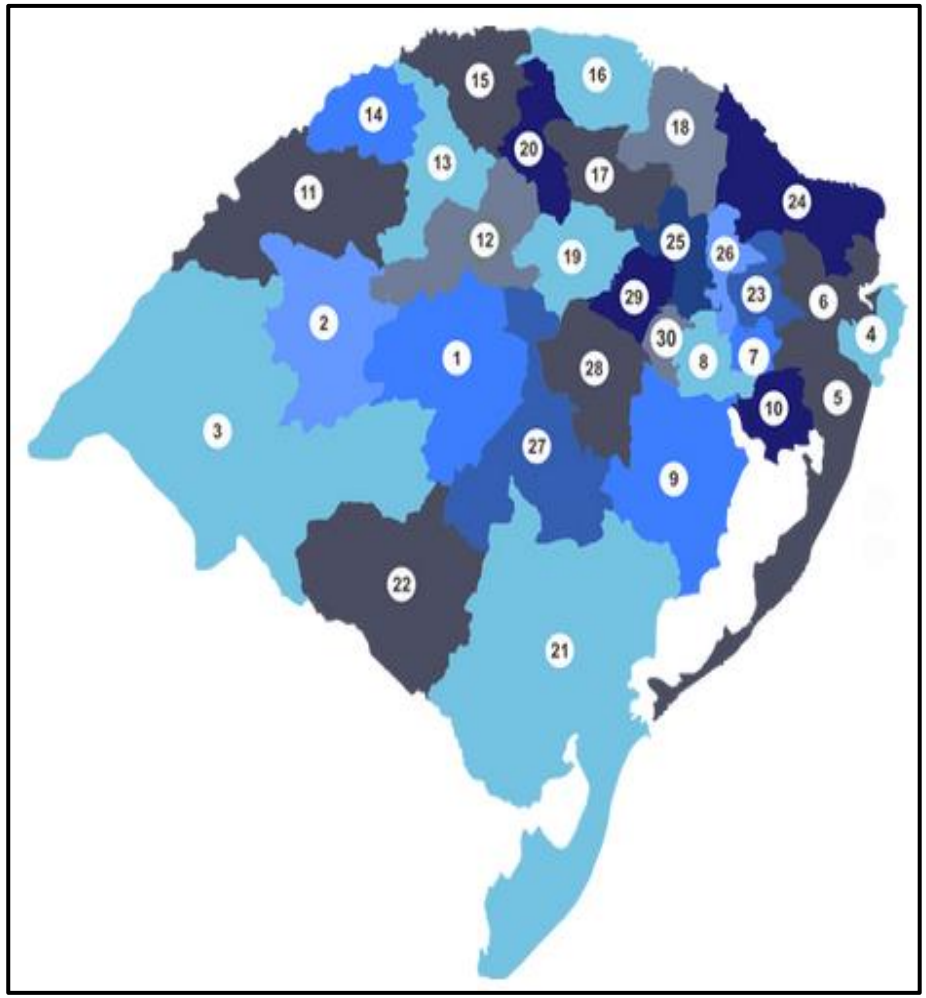






ANEXO A – Divisão político-administrativa da saúde pública no RS

As 30 Regiões de Saúde do RS



As 21 Regiões Covid