MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DE CIÊNCIAS DA SAÚDE DE PORTO ALEGRE PRÓ-REITORIA DE PESQUISA E PÓS-GRADUAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM TECNOLOGIAS DA INFORMAÇÃO E GESTÃO EM SAÚDE

Ursula Roséli Lamb

DESCOBERTA DE CONHECIMENTO EM BASE DE DADOS SOBRE PROCEDIMENTOS ASSISTENCIAIS EM UM PLANO DE SAÚDE

Porto Alegre
2023

Ursula Roséli Lamb

DESCOBERTA DE CONHECIMENTO EM BASE DE DADOS SOBRE PROCEDIMENTOS ASSISTENCIAIS EM UM PLANO DE SAÚDE

Dissertação no Programa de Mestrado Acadêmico em Tecnologias da Informação e Gestão em Saúde da Universidade Federal de Ciências da Saúde de Porto Alegre.

Orientador: Prof. Dr. Silvio César Cazella

Porto Alegre

2023

Catalogação na Publicação

Lamb, Ursula Roseli
DESCOBERTA DE CONHECIMENTO EM BASE DE DADOS SOBRE PROCEDIMENTOS ASSISTENCIAIS EM UM PLANO DE SAÚDE / Ursula Roseli Lamb. -- 2023.
106 f. : il., tab. ; 30 cm.

Dissertação (mestrado) -- Universidade Federal de Ciências da Saúde de Porto Alegre, Programa de Pós-Graduação em Tecnologias da Informação e Gestão em Saúde, 2023.

Orientador(a): Silvio César Cazella.

1. Mineração de dados. 2. Aprendizado de máquina não supervisionado. 3. Agrupamento de dados. 4. Planos de saúde. I. Título.

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

Ursula Roséli Lamb

DESCOBERTA DE CONHECIMENTO EM BASE DE DADOS SOBRE PROCEDIMENTOS ASSISTENCIAIS EM UM PLANO DE SAÚDE

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Tecnologias da Informação e Gestão em Saúde da Universidade Federal de Ciências da Saúde de Porto Alegre como requisito parcial para a obtenção do título de Mestre em Tecnologia da Informação e Gestão em Saúde

Orientador: Prof. Dr. Silvio César Cazella

Aprovada em:_____de___de___.

BANCA EXAMINADORA:

Prof. Dr.

Universidade

Prof. Dr.

Universidade

Prof. Dr.

Universidade

AGRADECIMENTOS

Agradeço a minha doce Lorena, que nasceu durante este programa de pós-graduação, e se tornou minha maior fonte de inspiração, de energia e de vontade de se tornar uma pessoa melhor.

Ao meu marido André Lazari pelo apoio.

A minha família, em especial a minha mãe e irmãs por cuidarem da Lorena, aos finais de semana, para que eu pudesse me dedicar à escrita deste trabalho.

Ao meu orientador Silvio Cazella não só pela compreensão em relação ao nascimento da Lorena e a minha dedicação ao mestrado, mas também pelos ensinamentos e pela condução docente.

Ao saudoso Dr. Rogério Dornelles e a minha amiga e parceira profissional Luciane La Rocca que junto comigo construíram a ideia inicial desta dissertação. À Luciane, agradeço também, pela paciência e entendimento em relação às ausências nas atividades profissionais para dedicação às aulas e aos estudos.

Ao Newton Lehugeur e Arthur Luiz Haubert pela confiança.

A Gabriela Toledo Costa, bolsista de iniciação científica, pela ajuda em relação aos materiais didáticos necessários à construção deste trabalho.

RESUMO

Introdução: O alto custo dos procedimentos que culmina em mensalidades elevadas dos planos de saúde e a necessidade constante de fomentar programas de cuidado em saúde aliada com a entrega ao paciente de saúde baseada em valor justificam novas abordagens em gestão e o uso de tecnologias no gerenciamento do sistema de saúde. Os procedimentos de saúde assistenciais realizados dentro do sistema de saúde suplementar formam bancos de dados com enormes oportunidades em gestão populacional. Técnicas de exploração de conjunto de dados como as disponíveis no processo de descoberta de conhecimento podem evidenciar conhecimento útil e novo para a gestão. **Objetivo:** Aplicar o processo de descoberta de conhecimento em bases de dados de procedimentos de saúde buscando extrair conhecimento útil e novo para apoiar a decisão na gestão a fim de melhorar a qualidade dos cuidados em saúde. **Método:** Esta pesquisa tem natureza aplicada, constituindo na exploração dos dados oriundos dos bancos de dados sobre procedimentos em saúde. Um modelo descritivo de aprendizado de máquina não supervisionado, seguindo a metodologia do processo de descoberta de conhecimento e bases de dados, foi elaborado a partir do algoritmo *K-means*. A quantidade de agrupamentos obtida foi orientada pelo índice de silhueta e o teor da segregação foi analisada de acordo com o conhecimento acerca do negócio e dos procedimentos. **Resultado:** O melhor modelo obtido permitiu que se descobrissem quatro agrupamentos. Esse modelo foi arquitetado nas etapas de pré-processamento. Foram excluídas, por exemplo, as variáveis que apresentaram pouca frequência de utilização assim como as que totalizaram valores baixos de sinistralidade. Também foram analisadas a faixa etária de utilização e os valores totais de gastos por códigos para que o algoritmo pudesse expressar seu melhor resultado. **Conclusões:** O processo de descoberta de conhecimento mostrou-se uma alternativa interessante para a exploração do *dataset* composto por procedimentos em saúde. Essa metodologia permitiu a identificação de um modelo descritivo que poderá auxiliar no processo de tomada de decisão em gestão de saúde. Ações em saúde relacionadas a grupos de portadores de doenças crônicas assim como à super utilizadores do plano de saúde, por exemplo, poderão ser melhor executadas a partir desses achados.

Palavras-chave: Mineração de dados. Aprendizado de máquina não supervisionado. Agrupamento de dados. *K-means*. Planos de saúde.

ABSTRACT

Introduction: The high cost of procedures that culminates in high monthly fees for health plans and the constant need to promote health care programs combined with the delivery of value-based health to the patient justify new approaches in management and the use of technologies in management of the health system. The health care procedures carried out within the supplementary health system form databases with enormous opportunities in population management. Data set exploration techniques such as those available in the knowledge discovery process can reveal useful and new knowledge for management. **Objective:** To apply the knowledge discovery process in databases of health procedures, seeking to extract useful and new knowledge to support management decisions in order to improve the quality of health care. **Method:** This research has an applied nature, constituting a single case study. The exploration of data from the databases on health procedures was carried out based on the process of discovering knowledge in the database. A descriptive model of unsupervised machine learning was elaborated from the K-means algorithm. The amount of clusters obtained was guided by the silhouette index and the content of segregation was analyzed according to knowledge about the business and procedures. **Result:** The best model obtained allowed four clusters to be discovered. This model was architected in the pre-processing steps. For example, variables that showed little frequency of use were excluded, as well as those that totaled low accident rates. The age group of use and the total values of expenses per code were also analyzed so that the algorithm could express its best result. **Conclusions:** The knowledge discovery process proved to be an interesting alternative for exploring the dataset composed of health procedures. This methodology allowed the identification of a descriptive model that could help in the decision-making process in health management. Health actions related to groups of people with chronic diseases as well as super users of the health plan, for example, could be better implemented based on these findings.

Keywords: Data mining. Unsupervised machine learning. Cluster analysis. K- means. Health plans.

LISTA DE QUADROS

Quadro 1 – Artigos relacionados e informações sobre tipo de dados e de aprendizado de máquina utilizados 33
Quadro 2 - Estrutura da base de procedimentos assistenciais em saúde provenientes do sistema de saúde suplementar 36
Quadro 3 - Quantitativo de instâncias e atributos 52
Quadro 4 - Etapas do Pré-processamento 52
Quadro 5 - Etapas do DCBD e atividades 56
Quadro 6 - Interpretação dos agrupamentos 84
Quadro 7 – Estatística descritiva – Valores em reais (R$) 84

LISTA DE FIGURAS

Figura 1 - Processo de Descoberta de Conhecimento em Base de Dados 22
Figura 2 - Funcionalidades do Orange em relação aos dados e pré-processamento 40
Figura 3 - Funcionalidades do Orange em relação às opções de visualização de resultados e em relação ao aprendizado de máquina supervisionado 40
Figura 4 - Funcionalidades do Orange em relação ao aprendizado de máquina não supervisionado 41
Figura 5 - Modelo de análise no Orange Data Mining 41
Figura 6 - Custo X Idade 47
Figura 7 - Distribuição de usuários por idade 47
Figura 8 - Idade e total de usuários com mais de 40 anos 48
Figura 9 - Informações sobre a base de dados 53
Figura 10 - Modelo de mineração de dados 53
Figura 11 - Agrupamentos 54
Figura 12 - Análise de componentes principais 55
Figura 13 - Representação da quantidade de usuários pela idade 58
Figura 14 - Representação da idade e custos 58
Figura 15 - Representação dos custos por sexo e idade 59
Figura 16 - Representação da quantidade de usuários por sexo 59
Figura 17 - Custo x códigos da base de dados utilizada 60
Figura 18 - Frequência de usuário por custo 61
Figura 19 - Avaliação Silhouette 61
Figura 20 - Dendograma com a representação dos agrupamentos 62
Figura 21 - Agrupamentos em escala multidimensional 63
Figura 22 - Representação dos agrupamentos pelo custo de utilização 64
Figura 23 - Representação dos agrupamentos pela idade dos usuários 64
Figura 24 - Representação dos agrupamentos pelo sexo dos usuários 65
Figura 25 - Variação dos custos por agrupamento por idade 65
Figura 26 - Internação Hospitalar por custo separada por agrupamento 66
Figura 27 - Terapias Oftalmologia por custo separada por agrupamento 66
Figura 28 - Terapias Quimioterapia por custo separada por agrupamento 67

Figura 29 - Internações Ortopedia e traumatologia por custo separada por agrupamento.....67
Figura 30 - Terapia ortopedia e traumatologia separada por agrupamento e valores.....68
Figura 31 - Principais variáveis atribuídas ao agrupamento 1.....68
Figura 32 - Internações Gastroenterologia por custo e agrupamento.....69
Figura 33 - Consultas médicas endocrinologia.....70
Figura 34 - Consultas médicas em Pronto Socorro.....70
Figura 35 - Terapias Ortopedia e Traumatologia.....71
Figura 36 - Consultas médicas Ginecologia e Obstetrícia.....71
Figura 37 - Consultas médicas Oftalmologia.....72
Figura 38 - Principais variáveis atribuídas ao agrupamento 2.....73
Figura 39 - Internações Hospital por custo e agrupamento.....74
Figura 40 - Internações Medicina Física e reabilitação.....75
Figura 41 Internações cardiologia.....75
Figura 42 - Internações clínica médica.....76
Figura 43 - Consultas médicas em pronto socorro.....76
Figura 44 - Exames serviços de imagem.....77
Figura 45 - Internações anestesiologia.....77
Figura 46 - Internações hospital.....78
Figura 47 - Principais variáveis atribuídas ao agrupamento 3.....79
Figura 48 - Consultas médicas Mastologia do agrupamento 3.....80
Figura 49 - Internações Oncologia Clínica do agrupamento 3.....81
Figura 50 - Terapias hospital do agrupamento 3.....81
Figura 51 - Principais variáveis atribuídas ao agrupamento 4.....82
Figura 52 - Terapias quimioterapia.....82
Figura 53 - Consultas psicologia.....83
Figura 54 - Modelo de pirâmide de risco.....85

LISTA DE ABREVIATURAS E SIGLAS

BI – Business Intelligence

OLAP – Online Analytical Processing

TISS – Troca de Informação em Saúde Suplementar ANS – Agência Nacional de Saúde Suplementar

DCBD – Descoberta de Conhecimento em Bases de Dados KDD – Knowledge Discovery in Databases

DM – Data Mining

EIS – Executive Information System

DSS – Decision system system

LGPD – Lei Geral de Proteção de Dados Pessoais SIP – Sistema de Informação de Produtos

PCA - Principal Components Analysis

MCA - Multiple Components Analysis

SUMÁRIO

1 INTRODUÇÃO 14
1.1 PERGUNTA DE PESQUISA 17
1.2 JUSTIFICATIVA 17
1.3 CONTRIBUIÇÃO 17
2 OBJETIVOS 20
2.1 OBJETIVO GERAL 20
2.2 OBJETIVOS ESPECÍFICOS 20
3 FUNDAMENTAÇÃO TEÓRICA 22
3.1 DESCOBERTA DE CONHECIMENTO EM BASE DE DADOS 22
3.1.1 Mineração de dados 23
3.1.1.1 Aprendizado de máquina 25
3.1.1.2 Gestão em saúde populacional 29
3.2 ATENÇÃO PRIMÁRIA EM SAÚDE 30
3.3 MEDICINA BASEADA EM VALOR 31
3.4 TRABALHOS RELACIONADOS 31
4 MATERIAIS E MÉTODOS 34
4.1 MATERIAIS 34
4.2 METODOLOGIA 36
4.3 FERRAMENTA 38
4.4 AGRUPAMENTO 42
4.5 ESCOLHA DO ALGORITMO 43
5 PROCESSO DE DESCOBERTA DE CONHECIMENTO EM BASE DE DADOS SOBRE PROCEDIMENTOS DE SAÚDE 45
5.1 PRÉ-PROCESSAMENTO 45
5.2 MINERAÇÃO DE DADOS 52
5.3 PÓS-PROCESSAMENTO 55
6 RESULTADOS E DISCUSSÃO 57
6.1 ANÁLISE CONJUNTA DOS DADOS E AGRUPAMENTOS 57
6.2 ANÁLISE DOS AGRUPAMENTOS 68
6.2.1 Agrupamento 1 – Pacientes com condições simples de saúde 68
6.2.2 Agrupamento 2 – Pacientes com altos custos hospitalares 73
6.2.3 Agrupamento 3 – Pacientes oncológicos com internações 79

6.2.4 Agrupamento 4 – Pacientes oncológicos 82
7 CONCLUSÃO 86
7.1 TRABALHOS FUTUROS 87
7.2 LIMITAÇÕES DA PESQUISA 88
REFERÊNCIAS 89
APÊNDICE A – CÓDIGOS SIP 94
APÊNDICE B – CÓDIGOS DE PROCEDIMENTO PRINCIPAL 95
ANEXO A - GUIA DE PUBLICAÇÃO DE ARTIGOS PARA AUTORES DA REVISTA HEALTH POLICY AND TECHNOLOGY 96

1 INTRODUÇÃO

O sistema de saúde suplementar enfrenta atualmente, assim como o setor público, os efeitos do aumento dos custos assistenciais. O incremento das despesas neste segmento, está associado ao envelhecimento da população, às mudanças no perfil de consumo dos produtos de saúde, à incorporação de novas tecnologias, à falta de eficiência das ações preventivas e aos modelos de remuneração. Essa problemática não é somente nacional e atinge o mundo inteiro. Poder prever esses custos e trabalhar a assistência, por meio de processos em cuidado em saúde, são algumas sugestões para tentar resolver esse problema (BERTSIMAS et al., 2008) e (FONSECA; OGATA, 2021).

Processos em cuidado em saúde são uma série de atividades destinadas a diagnosticar, tratar e prevenir doenças com a finalidade de melhorar a saúde das pessoas. Esses processos são aplicados por meio de soluções clínicas e não-clínicas e executados por diferentes profissionais dentro de uma equipe multidisciplinar. Além de aumentar a qualidade de vida dos pacientes, esses processos são responsáveis por reduzir o custo dos serviços, equalizar a demanda/oferta de atendimento em saúde, aumentar a produtividade dos profissionais envolvidos e a transparência na gestão dos custos (ROJAS et al., 2016).

Somado ao significado de processo em cuidado, como uma das maneiras de se melhorar os custos assistenciais, podemos citar o conceito de Triple Aim. Essa estrutura, estabelecida pelo Instituto de Melhoria da Saúde (*IHI – Institute for Healthcare Improvement*), caracteriza as diretrizes de gestão em saúde em três pilares a saber: melhoria da experiência do paciente, melhoria da saúde populacional e redução de custos (IHI, 2019). Conseguir trabalhar essas três estruturas por meio de análises de dados e melhorar o cuidado em saúde de uma população é uma boa alternativa para aprimorar o sistema de saúde tanto público quanto privado.

A gestão de saúde populacional, como área responsável em gerir ações em saúde, se propõe a projetar sistemas em torno de segmentos definidos da população, visando os indivíduos por meio da estratificação por risco de um evento adverso, como hospital, reinternação ou início de doença. A segmentação e a estratificação podem melhorar o atendimento ao paciente e gestão e informar o desenho de sistemas de cuidados. (YUILL; KUNZ, 2022).

As áreas de negócios usam dados para ganhar competitividade, aumentar eficiência e melhorar os serviços oferecidos aos clientes (FAYYAD et al., 1996). Ferramentas de análises de dados e BI (*business intelligence*) como armazenamento de dados, mineração de dados, processamento analítico online (OLAP – *online analytical processing*), dashboards e uso de sistemas baseados na nuvem para apoio a decisões são os pilares da gestão moderna (SHARDA et al., 2019). Logo, a tomada de decisão orientada por dados refere-se à prática de basear as decisões na análise dos dados, em vez de apenas na intuição. Logo essas ferramentas são essenciais na gestão populacional, também (PROVOST et al., 2016).

A Mineração de Dados é uma das tecnologias mais promissoras da atualidade uma vez que permite extrair conhecimento a partir de bases de dados. Ela, frequentemente, usa algoritmos de aprendizado de máquina para descobrir padrões em dados que podem destacar informações para melhorar o conhecimento sobre os clientes e dar suporte a melhores processos de decisão. Um dos fatores que reforçam a oportunidade no uso desta tecnologia, consiste na grande quantidade de dados atualmente disponíveis, ainda com pouca utilização efetiva. Ela surgiu, assim, como uma das soluções de sucesso para se analisar o conteúdo de dados digitais convertendo-os de meros dados acumulados e incompreensíveis em informações de valor cognitivo que os tomadores de decisão podem explorar e se beneficiar (DOS SANTOS, DIAS e FILHO, 2021; DA COSTA, CAZELLA e RIGO, 2014 e ABDELMAGID e QAHMASH, 2022).

O uso de técnicas de aprendizado de máquina, caracterizadas como uma das etapas da mineração de dados e do processo de descoberta de conhecimento, vem aumentando para prever resultados em saúde, incluindo custo, utilização e qualidade dos serviços ofertados. Essas técnicas têm sido usadas para predizer geradores de custos ou pacientes que passam a despender altos custos em saúde de forma rápida, além da possibilidade de ser usada na identificação de pacientes mais propensos a uma reinternação hospitalar (DOUPE et al., 2019).

Métodos de aprendizado de máquina chamados de não supervisionados são importantes ferramentas analíticas que podem facilitar a análise e a interpretação de dados de alta dimensão. Eles identificam padrões latentes e estruturas ocultas em ambientes de alta dimensão de dados e podem ajudar a simplificar conjuntos de dados complexos. Esse método pode ser incorporado às ciências da saúde como uma maneira de identificar fatores de risco para doenças, melhorar estratégias de

prevenção delas e facilitar a entrega de tratamentos baseados em medicina personalizada e contribuir no cuidado em saúde. Por conseguinte, as técnicas de aprendizado de máquina não supervisionado são uma ótima opção de ferramenta gerencial em saúde (ECKHARDT et al., 2022).

As empresas e associações, que oferecem planos de saúde aos seus colaboradores, recebem, mensalmente, relatório com todos os procedimentos de saúde realizados pelos seus associados/cooperados. Esses são enormes bancos de dados que podem servir de base de conhecimento de informação sobre a população referenciada. Assim, as operadoras de planos de saúde suplementar armazenam grande quantidade de dados referentes aos procedimentos realizados pelos beneficiários. O possível conhecimento existente em tal base pode auxiliar na tomada de decisões em programas de prevenção de doenças, caracterizar o perfil da população que utiliza o plano de saúde e ajudar no cuidado em saúde.

Portanto, diante da oportunidade de se analisar e de se entender o perfil de utilização de usuários de plano de saúde aliados à necessidade de se encontrar alternativas para melhorar a situação econômica e a sustentabilidade dessas organizações, a lacuna teórica da pesquisa é a aplicação de técnicas de aprendizado de máquina a bases de dados provenientes do sistema de saúde suplementar brasileiro.

Fez-se uma pesquisa na literatura científica para identificar estudos semelhantes e foram encontrados poucos artigos que abordam a temática do aprendizado de máquina não supervisionado testado e modelado a partir de dados provenientes de planos de saúde. Araújo, Santana e Santos Neto (2016) utizaram dados de planos de saúde para construir um modelo de aprendizado de máquina não suppervisionado para judar nas autorizações de procedimentos de saude. Artigos com temática semelhante e que utilizaram dados de plano de saúde (Brasil) ou seguro saúde (outros países) como o de Xie, Schreier, Hoy, Liu, Neubauer, Chang, Redmond e Lovell (2016), Seleme, Cubas e Carvalho (2023), Zhang, Li e McConnell (2021) e Araújo, Santana e Santos Neto (2016) foram compreendidos, mas nenhum deles utilizou o aprendizado de máquina não supervisionado. Essa classificação de aprendizado de máquina, porém, foi utilizada nos estudos de Soleymani, Yaseri, Farzadfar, Mohammadpour, Sharifi e Kabir (2018), Santos, Dias e Chiavegatto Filho (2021)e Forkan, Khalil e Kumarage (2020), por exemplo, mas foram utilizados os mais diversos dados em saúde, como questionários epidemiológicos, dados de sinais vitas

e de prescrições médicas. Assim, diante das particularidades do sistema de saúde suplementar operado pelos planos de saúde em relação aos Estados Unidos, por exemplo, que praticam a modalidade de seguro saúde, este estudo contribuirá com a oferta de materiais científicos que tratam do assunto de inteligência artificial e planos de saúde.

1.1 PERGUNTA DE PESQUISA

A questão de pesquisa deste estudo é: “Como podemos utilizar o processo de descoberta de conhecimento em base de dados para caracterizar usuários de planos de saúde de acordo com a sua utilização e, a partir dessas descobertas, possibilitar a melhoria da sua condição física e mental por meio de ações em gestão de saúde populacional?”

1.2 JUSTIFICATIVA

Esta pesquisa surge de uma necessidade real para áreas de negócio que precisam analisar carteiras de beneficiários de planos de saúde. A área apresenta necessidade de aplicação de soluções que possam auxiliar na tomada de decisão para apoiar gestores quanto às características de uso de procedimentos em saúde. Com o acúmulo e produção extensa de dados sobre os associados entende-se que a extração de informações com base em relatórios já não supre mais as necessidades para planejamento futuro. Sendo assim, assume-se que a mineração de dados por meio da aplicação algoritmos de aprendizado de máquina não supervisionado poderá apresentar contribuições aos gestores de saúde. Essas contribuições se referem à descoberta de características de utilização dos usuários do sistema de saúde suplementar e como eles podem ser segmentados a fim de direcionar ações de melhoria em saúde.

1.3 CONTRIBUIÇÃO

Os resultados das análises, realizadas com o uso das técnicas de mineração de dados, deverão gerar insights para a tomada de decisão em ações de melhoria na saúde, como preventivas e no acompanhamento de doentes crônicos, por exemplo.

Ainda, há a possibilidade de se identificar quais populações correspondem à maior sinistralidade nos planos de saúde e assim colaborar para o equilíbrio econômico-financeiro das carteiras. Por conseguinte, a descoberta de conhecimento em base de dados oriundos do sistema de saúde suplementar se mostra importante e relevante por ser um meio de se caracterizar os usuários dos planos de saúde e de dar condições à oferta de programas de políticas de melhoria e de cuidado em saúde a eles. Por conseguinte, conhecer a saúde dessas comunidades e sua evolução é crucial não apenas para a avaliação das metas estabelecidas, mas também para encontrar estratégias para aprimorar a saúde dos membros como um todo (SHARDA et al., 2019).

Esta dissertação está organizada da seguinte forma: No capítulo 1 apresenta-se a introdução. Nela está contida informações acerca do contexto teórico em que o tema deste trabalho está inserido assim como os subcapítulos que informam à qual pergunta de pesquisa este estudo se refere. Também constam a justificativa e a contribuição deste estudo para a comunidade.

No capítulo 2 são demonstrados os objetivos tanto o geral quanto os específicos.

No capítulo 3 está descrita a fundamentação teórica, que está subdividida em descoberta de conhecimento em bases de dados, gestão em saúde populacional, atenção primária em saúde, medicina baseada em valor e trabalhos relacionados. Ele está organizado desta maneira, pois o método de pesquisa utilizado é o processo de descoberta de conhecimento em base de dados. As áreas de gestão de saúde populacional, de atenção primária em saúde e medicina baseada em valor foram incluídas neste trabalho pois servem de guia para ações de melhoria em saúde que podem ser tomadas a partir do conhecimento adquirido com o processo de DCBD. Neste capítulo, ainda, foram descritos alguns artigos sobre mineração de dados com algoritmos de agrupamento que serviram de base para esta dissertação.

Já no capítulo 4, apresenta-se os materiais e métodos. Nele constam as informações de como a base de dados foi escolhida e adaptada para que o algoritmo apresentasse os melhores resultados vinculados à pergunta de pesquisa. São expostas a ferramenta de mineração de dados utilizada e a fundamentação do algoritmo escolhido.

No capítulo 5 é abordado o processo de descoberta de conhecimento em base de dados sobre procedimentos de saúde. Ele foi divido em pré- processamento,

mineração de dados e pós-processamento como uma forma de melhorar a didática de apresentação dos assuntos. No pré-processamento foram incluídas as etapas do DCBD chamadas de seleção, limpeza e transformação dos dados. Na mineração de dados constam as etapas de mineração de dados e no pós-processamento a informação de como o processo de conhecimento e aplicação prática dos resultados obtidos na mineração foram feitos.

No capítulo 6 tem-se os resultados e a discussão. Nele, os resultados obtidos na mineração de dados e no pós-processamento são discutidos e apresentados. A discussão baseia-se, também, nas áreas de conhecimento apresentadas no capítulo 3. Os resultados, como capítulo foi dividido com base nas análises dos cluster de forma conjunta e individual.

Esta dissertação finaliza no capítulo 7 em que são apresentadas as conclusões deste estudo com informações acerca de trabalhos que poderão ser realizados, futuramente, a partir dos achados deste estudo e suas limitações.

2 OBJETIVOS

2.1 OBJETIVO GERAL

O objetivo geral desta pesquisa é explorar a base de dados sobre procedimentos em saúde provenientes da saúde suplementar por meio da aplicação do processo de descoberta de conhecimento em base de dados, visando extrair conhecimento para apoio ao processo de tomada de decisão.

2.2 OBJETIVOS ESPECÍFICOS

Os objetivos específicos são:

- Analisar as bases de dados provenientes dos procedimentos em saúde a fim de iniciar um conhecimento sobre procedimentos utilizados e gasto total de cada usuário;
- Pré-processar os dados;
- Aplicar a mineração de dados por meio do aprendizado de máquina não supervisionado;
- Descrever os agrupamentos de dados obtidos apresentando os padrões encontrados;
- Descobrir perfis de utilização;
- Avaliar o conhecimento extraído;

O conhecimento extraído será avaliado de acordo com a análise e o conhecimento que se tem sobre a base de dados e os procedimentos dos planos de saúde. Ou seja, espera-se encontrar grupos de pessoas que utilizam procedimentos semelhantes e que possam ser caracterizados de alguma forma dentro de uma sistemática de uso do sistema de saúde suplementar. Esses grupos serão avaliados e selecionados de acordo com a sua relevância no contexto da saúde populacional.

A análise de bancos de dados de procedimentos assistenciais provenientes da saúde suplementar pode contribuir para a gestão de saúde populacional de maneira a manejar grupos de ofensores de plano de saúde, ou seja de usuários que apresentam alto consumo de serviços de saúde, e contribuir para uma gestão econômico-financeira saudável da carteira além de melhorar o bem-estar desses beneficiários por meio de ações em saúde e de utilização dos serviços dentro do

sistema médico-hospitalar.

3 FUNDAMENTAÇÃO TEÓRICA

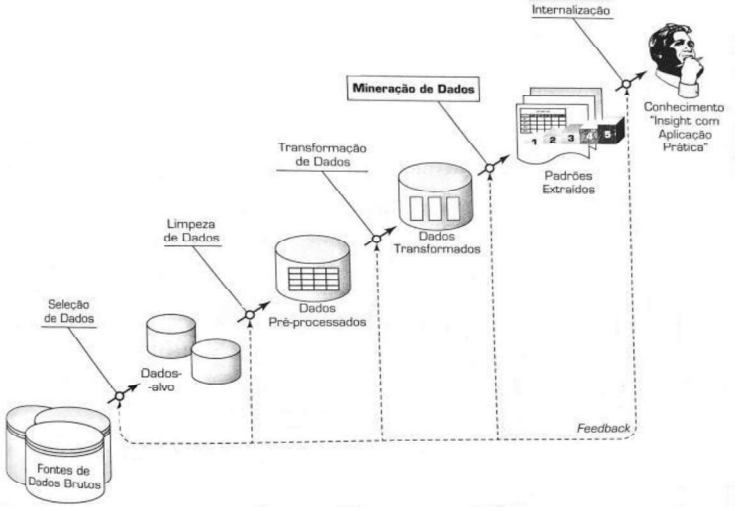
3.1 DESCOBERTA DE CONHECIMENTO EM BASE DE DADOS

Devido à grande quantidade de dados que as empresas e instituições de saúde coletam e armazenam se faz necessário o uso de teorias e de ferramentas computacionais para a extração de conhecimento útil desses dados. Esses métodos de conhecimentos aplicáveis em bancos de dados deram origem ao que chamamos de Descoberta de Conhecimento em Banco de Dados (DCBD), do inglês, *Knowledge Discovery in databases* (KDD).

Nos últimos anos, novas abordagens foram incluídas neste processo, incluindo métodos para pré-processar, integrar, analisar e interpretar dados biomédicos com o objetivo de identificar hipóteses testáveis. Além disso, é reconhecida a necessidade de se incluir o usuário final no processo interativo de descoberta de conhecimento com o objetivo de apoiar a inteligência humana com a inteligência da máquina (HASSANI-PAK; RAWLINGS, 2017).

O processo de descoberta de conhecimento em base de dados está ilustrado na figura 1 e pode ser definido como um processo que utiliza métodos de mineração de dados para encontrar informações e padrões úteis nos dados. As suas etapas são descritas como seleção de dados, pré-processamento de dados, transformação de dados, mineração de dados e interpretação/avaliação de dados (SHARDA et al., 2019).

Figura 1 - Processo de Descoberta de Conhecimento em Base de Dados



Fonte: Sharda et al. (2019).

Neste contexto faz-se necessário citar o que se tem chamado de “Big Data”. Esse termo refere-se a grandes bancos de dados ou a grandes quantidades de dados que não podem ser analisados, pesquisados, armazenados e interpretados com o uso de técnicas estatísticas tradicionais de pré-processamento. São conjuntos de dados de saúde tão grandes e complexos que são difíceis de gerenciar com software e/ou hardware rotineiros (NETTO et al., 2022). Como exemplos de grandes bancos de dados podemos citar as informações coletadas de celulares, de mídias sociais, de dados sociodemográficos, de dados genômicos e de dados de prontuários de saúde. Eles são a base do processo de DCBD (KRITTANAWONG et al., 2017).

Análises de grandes bancos de dados em medicina e saúde integram a compreensão multidisciplinar de áreas como a de bioinformática, imagens médicas, sensores de sinais e de dados médicos e de saúde. Elas estão aptas a investigar grande quantidade de informações de pacientes identificando associações e correlações entre elas, assim como desenvolver modelos preditivos usando técnicas de mineração de dados. As análises abrangem a integração de dados heterogêneos, controle da qualidade dos dados, modelagem, interpretação e validação da informação. O conhecimento adquirido nesses processos fornece benefícios para os pacientes, para os agentes envolvidos com assistência em saúde e na formulação de políticas públicas (RISTEVSKI e CHEN, 2018).

3.1.1 Mineração de dados

A técnica da mineração de dados – *Data Mining* (DM) – é uma das alternativas mais eficazes para extrair conhecimento a partir de grandes volumes de dados, descobrindo relações ocultas, padrões e gerando regras para predizer e correlacionar dados, que podem ajudar as instituições nas tomadas de decisões mais rápidas ou, até mesmo, a atingir maior grau de confiança (GALVÃO et al., 2009). Ela se refere a um conceito relativamente novo que foi introduzido em meados da década de 1990 como uma nova abordagem para análise de dados e descoberta de conhecimento. É o processo de descobrir automaticamente informações úteis em grandes repositórios de dados. Sendo um domínio altamente orientado a aplicativos, a mineração de dados incorporou muitas técnicas de outros domínios, como estatísticas, aprendizado de máquina, reconhecimento de padrões, banco de dados e sistemas de armazenamento de dados, recuperação de informações, visualização, algoritmos, computação de alto

desempenho e muitos domínios de aplicativos (SOLEYMANI et al., 2018).

Pode ser explicada, também, como a aplicação de algoritmos específicos para extração de padrões dos dados. Por conseguinte, essa área do conhecimento se encarrega em desenvolver métodos e técnicas que servem para dar sentido aos dados coletados (FAYYAD et al., 1996). É um processo não trivial de identificar padrões válidos, novos, potencialmente úteis e posteriormente compreensíveis junto à base de dados (SHARDA et al., 2019; FAYYAD et al., 1996). Entre suas várias tarefas, destacam-se algumas que são as mais utilizadas: associação, classificação, regressão, agrupamento e sumarização. (GALVÃO et al., 2009; PROVOST et al., 2016).

O objetivo da mineração de dados, enquanto etapa do processo DCBD, é descobrir, de forma automática ou semiautomática, o conhecimento disponível em grandes massas de dados armazenadas nos bancos de dados das organizações, permitindo, desta forma, agilidade na tomada de decisão. Uma organização que emprega o processo de DCBD na análise de seus dados é capaz de: 1) criar parâmetros para entender o comportamento dos dados, os quais podem ser referentes a pessoas envolvidas com a organização; 2) identificar afinidades entre dados que podem ser, por exemplo, entre pessoas e produtos e ou serviços; 3) prever hábitos ou comportamentos das pessoas e analisar hábitos para se detectar comportamentos fora do padrão entre outros (DA COSTA, CAZELLA e RIGO, 2014).

A utilização das técnicas de mineração de dados no contexto da descoberta de conhecimento pode ajudar pagadores terceirizados, como organizações de seguros de saúde, a extrair conhecimento útil de milhares de procedimentos e identificar subconjuntos deles para avaliação e análise mais aprofundada por fraude e abuso. Desta forma, a abordagem de mineração de dados faz parte de uma abordagem mais eficiente de um sistema de auditoria baseado em tecnologia da informação e eficácia (JOUDAKI et al., 2015).

Além de melhorar a administração de centros de cuidado em saúde, benefícios adicionais podem ser obtidos por meio dos processos de aplicação de mineração de dados em dados de saúde. Exemplifica-se que ela pode ajudar a entender o comportamento dos recursos e suas aplicações nos pacientes, a apresentar sugestões para redesenhar o processo de cuidado, a ajudar a obter “insights” e melhorar a colaboração entre pares, a prever o comportamento de pacientes de acordo com casos anteriores (ROJAS et al., 2016).

Os sistemas de informação executiva (EIS - *executive information system*) são sistemas especializados de apoio à decisão de informação (DSS - *decision system support*) que fornecem à alta administração empresarial apoio técnico para a tomada de decisões. Esses sistemas estão se tornando ferramentas diárias de gestão de negócio, tanto na área pública quanto na privada, devido à necessidade de atitudes rápidas e melhores em ambientes turbulentos e competitivos de negócios. A qualidade e a amplitude dos dados são cruciais para o entendimento da análise a ser feita e se ela poderá ser utilizada nos problemas de negócio e na extrapolação de tendências. Assim, técnicas de mineração de dados são consideradas a ligação entre conhecimento oculto do passado e o conhecimento esperado do futuro. Elas fazem o link entre tecnologia e conhecimento. Nestas técnicas são concentrados vários métodos para alcançar a geração de conhecimento usando a tecnologia como ferramenta e melhorando o gerenciamento de negócios (GLOVER et al., 2010).

3.1.1.1 Aprendizado de máquina

O aprendizado de máquina (*machine learning, do inglês*) é uma área da inteligência artificial e que pode ser utilizada dentro do processo que chamamos de mineração de dados. Ele representa técnicas usadas para resolver problemas com uso de bancos de dados. Essas técnicas utilizam mecanismos computacionais para identificar padrões de interação entre variáveis. Em contraste com as técnicas estatísticas tradicionais, o aprendizado de máquina tem como foco construir sistemas automatizados de decisão. O fundamento do aprendizado de máquina está inserido nas etapas de introduzir algoritmos em dados, aplicar análise computacional para predizer resultados dentro de valores aceitáveis de acurácia, identificar padrões e tendências nesses dados e aprender com todos esses processos (HANDELMAN et al., 2018).

A principal premissa do aprendizado de máquina é introduzir dados a algoritmos que são processados por meio de análises computacionais para predizer valores aceitáveis em intervalos de acurácia identificando padrões e ensinando o computador a repetir a operação com a mesma assertividade (Ibid.). Dependendo da forma como os padrões são extraídos dos dados históricos, os algoritmos de aprendizado dos métodos de mineração de dados podem ser classificados como supervisionados e não supervisionados. Ela depende dos tipos de dados que estão

sendo minerados, do conhecimento que se quer adquirir e dos tipos de algoritmos a serem utilizados. No caso dos algoritmos de aprendizado supervisionado, os dados de treinamento incluem tanto os atributos descritivos (variáveis independentes ou decisórias) quanto o atributo de classe (variável de saída ou de resultado). Ele tenta descobrir a relação entre variáveis de entrada (atributos) e uma variável de saída (atributo de destino ou variável dependente). São usados como finalidade de classificação e de predição (BISHARA; MAZE; MAZE, 2022).

Em contraste, no caso do aprendizado não supervisionado, os dados de treinamento incluem apenas os atributos descritivos e o objetivo é descrever associações e padrões entre essas variáveis. Não há informações de variáveis dependentes e variável resposta. Dentre as técnicas mais utilizadas nesta situação há a análise de componentes principais, análise de componentes independentes e a análise de conglomerados (agrupamentos) (MORETTIN et al., 2020; SHARDA et al., 2019; KRITTANAWONG et al., 2017). É utilizado num contexto de descrição dos padrões dos dados incluindo regras de associação e segmentação na detecção de anomalias e agrupamento (JOUDAKI et al., 2015).

3.1.1.1.1 Aprendizado de máquina não supervisionado

Tarefas de mineração de dados podem fazer uso de algoritmos de aprendizado de máquina, que realizam operações relacionadas a métodos de análises não supervisionadas. Esses algoritmos podem ser aplicados a banco de dados sem rótulos a fim de se encontrar agrupamentos de populações semelhantes. Eles extraem significado dos dados sem treinar um modelo de dados rotulados. Nesse modelo não há distinção entre variáveis responsivas e variáveis preditoras (PROVOST et al., 2016 e BRUCE e BRUCE, 2019).

No aprendizado de máquina não supervisionado, algoritmos computacionais analisam dados não classificados para reconhecer e determinar padrões e, posteriormente, esses padrões são avaliados por especialistas da área de negócio com o intuito de se comprovar que o modelo de análise de dados proposto será útil na tomada de decisão (HANDELAMN et al., 2018).

O agrupamento ou cluster - do inglês - formado pela aplicação do algoritmo *K-means* e o agrupamento hierárquico são dois exemplos de métodos deste tipo de aprendizado de máquina. Eles particionam dados em grupos com base em suas

semelhanças (ECKHARDT et al., 2022).

Na clusterização ou agrupamento, há a identificação de grupos nos dados. O algoritmo recebe os dados, analisa-os e determina quais semelhanças permitem que as características dos dados sejam agrupadas em subseções e padrões dentro dos dados a serem descobertos. Ela funciona maximizando as variações interclasses e minimizando as variações interclasses. Os grupos formados são descritos em termos de variáveis e categorias (HANDELMAN et al., 2018; BERTSIMAS et al., 2008).

Um agrupamento (ou cluster) é uma coleção de registros similares entre si, porém, diferentes dos outros registros nos demais agrupamentos. Em geral, as medidas de similaridade usadas são as medidas de distâncias tradicionais (Euclidiana, Manhattan etc.). Os elementos de um agrupamento são considerados similares aos elementos no mesmo agrupamento e dissimilares aos elementos nos outros agrupamentos (CAMILO; DA SILVA, 2009). De um modo geral são usados algoritmos k-means e o método hierárquico de classificação (GLOVER et al., 2013).

Esse tipo de aprendizado computacional pode simplificar e organizar dados complexos e pode ajudar pesquisadores a categorizar pessoas ou instituições. Como exemplo disso podemos citar os modelos que usam técnicas chamadas de análise de correspondência múltipla, do inglês *multiple componentes analysis* (MCA) e de análise de componentes principais, do inglês *principal componentes analysis* (PCA). Ambas podem ser usadas para redução de dimensionalidade. A PCA projeta pontos de dados em um espaço dimensional euclidiano para fornecer uma melhor representação de dados. As dimensões explicam a maior quantidade de variabilidade nos dados. Ela transforma dados de alta dimensão em funções lineares que explicam a variação total dos dados. O usuário deve padronizar e dimensionar os dados inseridos e decidir quantos componentes principais usar em análises subsequentes (DOS SANTOS, DIAS e FILHO, 2021 e ECKHARDT et al., 2022).

Os modelos de aprendizado de máquina não supervisionado podem ser aplicados com o objetivo de se atingir diferentes objetivos, como exemplos podemos citar:

- Criar uma regra preditiva na ausência de uma resposta rotulada;

- Identificar grupos de dados significativos por meio dos métodos de agrupamento;

- Reduzir a dimensão dos dados para um conjunto mais gerenciável de variáveis (esse conjunto poderia ser usado como modelo de regressão ou de

classificação). Por meio da redução dos dados para um conjunto menor de características pode-se construir um modelo mais potente e interpretável;

- Pode ser utilizado como uma extensão da análise exploratória dos dados em situações em que há uma grande quantidade de variáveis e registros. O objetivo é obter uma melhor percepção interna do conjunto de dados e de como as diferentes variáveis de relacionam umas com as outras.

As técnicas não supervisionadas de aprendizado de máquina oferecem meios de se filtrar e de se analisar variáveis e descobrir relacionamentos entre os dados. Também, é importante como um bloco de construção para técnicas de regressão e de classificação. Com grandes bancos de dados, se uma pequena subpopulação não for bem representada na população global, o modelo treinado pode não ter um bom desempenho naquela subpopulação. Como o agrupamento, é possível identificar e rotular subpopulações ou então a subpopulação pode ser representada com sua própria característica, forçando o modelo global a considerar explicitamente a identidade da subpopulação como uma preditora (BRUCE e BRUCE, 2019).

Esta ferramenta é especialmente importante para o “problema de partida a frio”. Nestes tipos de problemas, como na identificação de potenciais tipos de fraude, podemos inicialmente não ter nenhuma resposta para treinar o modelo. Com o tempo, conforme os dados são coletados, podemos aprender mais sobre o sistema e construir um modelo preditivo tradicional. Assim, o agrupamento nos ajuda a iniciar o processo de aprendizado mais rapidamente por meio da identificação de segmentos populacionais (BRUCE; BRUCE 2019).

As técnicas de agrupamento podem ser divididas em métodos de particionamento, hierárquicos, métodos baseados na densidade e métodos baseados em grade. No método de particionamento um conjunto D de dados com n registros e “k” o número de agrupamentos desejados, os algoritmos de particionamento organizam os objetos em k agrupamentos, tal que k <= n. Os algoritmos mais comuns de agrupamento são: *k-Means* e *k-Medoids*. Já no método hierárquico, cria-se o agrupamento por meio da aglomeração ou da divisão dos elementos do conjunto. A forma gerada por este método é um dendograma. Este método pode ser dividido em aglomerativo e divisivo (CAMILO; DA SILVA, 2009).

Os algoritmos de agrupamento têm sido amplamente utilizados em pesquisas biomédicas devido à alta capacidade de geração de dados nessa área. Eles são usados em processamento de linguagem natural biomédica, em ontologias, em

análises de imagens médicas e em dados fisiológicos, por exemplo. Também estão sendo usados nas análises de progressão de doenças, de surtos e agrupamentos de doenças relacionadas. Além disso, estão sendo usados como uma variável categórica em lugar de dados de biomarcadores de alta dimensão para determinar se agrupamentos estão associados com os principais resultados em saúde (LIM et al., 2017 e ECKHARDT et al., 2022).

O método hierárquico é uma abordagem alternativa de união que gera uma hierarquia de agrupamento aninhada que pode derivar em várias soluções de agrupamento. A homogeneidade do agrupamento é caracterizada pela inércia interna. O agrupamento hierárquico tem a vantagem de não precisar definir um número de agrupamentos a priori quando comparado ao agrupamento de particionamento como o *K-means*. Ele funciona juntando, gradualmente, os indivíduos mais semelhantes e é representado por um dendrograma (DOS SANTOS, DIAS e FILHO, 2021 e ECKHARDT et al., 2022).

O algoritmo chamado *K-means* particiona os dados em grupos distintos e não sobrepostos. Para o seu funcionamento há a necessidade de se organizar os dados sem rótulos e escolher os números de grupos que deverão ser formados (ECKHARDT et. al, 2022).

De acordo com Handelman et al (2018), na análise de variáveis de pacientes por meio de algoritmos de aprendizado de máquina não supervisionado foi possível identificar vários subgrupos de pacientes que têm diferentes desfechos clínicos embora possuíssem o mesmo diagnóstico.

O desempenho dessas técnicas não pode ser analisado pela acurácia, por isso a expertise dos pesquisadores se faz necessária. Assim, a validade da solução de agrupamento depende da experiência clínica e da utilidade da solução encontrada (DOUPE et al., 2019 e ECKHARDT et al., 2022).

3.1.1.2 Gestão em saúde populacional

Atualmente, os estudos e previsões relacionados à carga global de doenças, seus impactos e relações com fatores de risco, associados à necessidade do cuidado contínuo das doenças crônicas, vão ao encontro aos quatro objetivos principais da gestão em saúde: melhorar a experiência dos pacientes com os serviços; aprimorar a saúde da população; reduzir os custos e aperfeiçoar a experiência dos profissionais

deste segmento. Por conseguinte, ações em saúde que visam promover os objetivos de gestão e melhorar a experiência do paciente são de grande valia para o sistema de saúde, seja ele público ou privado (FONSECA; OGATA, 2021).

O gerenciamento populacional de saúde é adotado para melhorar os resultados para os indivíduos, personalizando os serviços para abordar sua saúde e necessidades de cuidados de uma forma que reconheça que a saúde é determinada tanto por fatores relacionados à programas preventivos e a por fatores socioeconômicos que pela própria prestação de cuidados de saúde (YUILL; KUNZ, 2022).

As doenças crônicas têm grande impacto sobre as causas de morte e causam grandes restrições às atividades de lazer e ocupações. As operadoras não têm conhecimento sobre a saúde dos seus beneficiários e a utilização dos serviços. Somente observa-se a existência de informações sobre questões financeiras relacionadas às despesas com os procedimentos. Desse modo, é fundamental que as empresas de saúde privada utilizem os recursos disponíveis para gerenciar a saúde dos seus usuários (FONSECA; OGATA, 2021).

O sistema de saúde suplementar é complexo e implica na necessidade de diversas estratégias corporativas a fim de manter a universalização do direito social à saúde e encontrar dificuldades em enfrentar custos dispendiosos operacionais. Logo, soluções baseadas em tecnologia são uma opção para combater esses impasses (FONSECA; OGATA, 2021).

3.2 ATENÇÃO PRIMÁRIA EM SAÚDE

A Atenção Primária à Saúde (APS) vem se apresentando como a melhor forma de organização dos serviços de saúde. Como benefícios associados aos programas que utilizam suas características tem-se uma maior participação comunitária nos cuidados da própria saúde; redução de mortalidade e de doenças relacionadas à ausência de condições sanitárias e econômicas; e o desenvolvimento de medidas preventivas e de promoção da saúde (PIRES et. al, 2019).

Ela pode ser conceituada como uma política em saúde com vistas a organizar os serviços para atender as necessidades da população (PIRES et. al, 2019).

A referência das práticas de saúde nos últimos anos demonstra assistência fragmentada, direcionada ao uso excessivo de procedimentos e especialidades,

cuidado desintegrado e descentralizado. Estes padrões resultam em ineficiência e desfechos desfavoráveis, pois são atrelados à demanda espontânea e utilização inapropriada. Nesse contexto, a Assistência Primária à Saúde (APS) está sendo considerada pelas operadoras, pois, vem contribuir para a maior equidade nos serviços. Traz a possibilidade de reduzir diferenças entre grupos populacionais, promovendo o cuidado coordenado, individualizado, focado na pessoa e não na doença, e integrado com as necessidades de saúde, desde a prevenção até a reabilitação. Esta estrutura traz maior satisfação dos clientes, porém, exige adaptação conforme o perfil heterogêneo da carteira, baseada nas informações referentes à população envolvida (FONSECA; FOGATA, 2021).

3.3 MEDICINA BASEADA EM VALOR

Quando se cita a medicina baseada em valor e o cuidado centrado no paciente há a referência em abordagens que buscam entregar resultados satisfatórios aos pacientes com a utilização eficiente dos recursos disponíveis.

Desenhar modelos de serviços em saúde de acordo com características biofísicas, relacionadas com o estilo de vida, ao aspecto social, aos fatores ambientais dos pacientes assim como considerar suas necessidades, a de seus familiares e a da sua comunidade desafiam os métodos gerenciais tradicionais que prestam serviços em saúde de forma fragmentada (SCHULTE; BOHNET- JOSCHKO, 2022).

Para a Organização Mundial de Saúde, os modelos de cuidado em saúde ditos centrados no paciente, não somente cuidam do paciente de forma integrada (sejam na detecção de doenças e cuidados agudos, crônicos e paliativos), mas se preocupam, também, com ações de prevenção e de promoção em saúde (SCHULTE; BOHNET-JOSCHKO, 2022).

Como exemplos de maneiras de se melhorar a experiência do paciente e de se entregar valor em saúde podemos citar o fortalecimento de modelos assistenciais focados em cuidados primários, criação de planos de saúde personalizados, acessos facilitados a registros de saúde nas diversas estruturas que compõem o sistema de saúde, ações que avaliam a experiência do paciente e o desempenho do profissional de saúde, ações voltadas à coordenação do cuidado e à encaminhamentos corretos dentro do sistema de saúde, dentre outras (SCHULTE; BOHNET-JOSCHKO, 2022).

3.4 TRABALHOS RELACIONADOS

Santos, Dias e Chiavegatto Filho (2021) utilizaram o algoritmo *k-means* para agrupar indivíduos que não possuem seguro saúde em Portugal. A base de dados utilizada consistiu nos formulários de saúde preenchidos pela população deste país. Eles obtiveram sucesso ao conseguir caracterizar essa população em três agrupamentos. Essa segmentação possui valor para o sistema de saúde do país ao identificar quem é essa população e como eles poderão ser beneficiados por ações públicas de saúde.

Meng et al. (2023), também utilizaram o algoritmo *k-means* para caracterizar em grupos quem são as pessoas que não tomaram a vacina para a COVID-19 nos Estados Unidos da América. A base de dados por eles utilizada foi composta pelos formulários BeSD (*Behavioral and Social Drivers of vaccionation*). Esse tipo de pesquisa foi criado pela Organização Mundial da Saúde para tentar entender o comportamento das pessoas anti-vacina.

Eles conseguiram agrupar essa população não vacinada, de acordo com suas peculiaridades, em três grupos. E essa divisão tem por finalidade ajudar os tomadores de decisão em saúde a explorar possíveis intervenções e políticas públicas.

Dados de seguro de saúde foram utilizados no estudo de Xie, Schreier, Hoy, Liu, Neubauer, Chang, Redmond e Lovell (2016) na produção de um modelo de aprendizado de máquina supervisionado para predizer dias de hospitlização.

Já, Soleymani, Yaseri, Farzadfar, Mohammadpour, Sharifi e Kabir (2018), utilizaram um modelo de aprendizado de máquina não supervisionado para detectar possíveis fraudes em prescrições médicas. Esse estudo utilizou dados de prescrições médicas do ano de 2013 realizadas no sistema de saúde privado do Irã. Os resultados mostraram que esse modelo pode ser usado para ajudar peritos na detecção de fraudes ao sistema de saúde. Forkan, Khalil e Kumarage (2020), também utilizaram o recurso do aprendizado de máquida não supervisionado, porém com dados correlacionados de sinais vitais múltiplos para simplificar a tarefa dos profissionais de saúde na tomada de decisões clínicas.

Dados de planos de saúde brasileiros foram utilizados nos estudos de Seleme, Cubas e Carvalho (2023) e de Araújo, Santana e Santos Neto (2016). Enquanto o primeiro fez uso de quantidades de consultas eletivas, quantidades de consultas em pronto-socorro, quantidade de internações, de exames, de terapias e de custo total de sinistralidade além de informações sobre questões de saúde com o intuito de se tentar

identificar variáveis relacionadas ao alto custo em saúde e à saúde mental; o segundo, desenvolveu um modelo de aprendizado de máquinda não supervisionado para apoiar os profissionais responsáveis pelas autorizações de procedimentos em saúde. Para isso utilizaram variáveis provenientes das guias de pré-autorização e de outras julgadas relevantes pelos profissionais experientes em autorizações a fim de se obter um mecanismo de suporte à decisão.

Zhang, Li e McConnell (2021) utilizaram dados do CID (classificação internacional de doenças) e de custos de plano de saúde de beneficiários relacionados à cuidados paliativos com a finalidade de se produzir um pipeline que ajude na predição de pacientes que passarão por esse processo e na ajuda em reduzir custos. Segundo essa publicação, o custo despendido nesta assistência no final da vida corresponde a uma média de 25% de todo o custo de um usuário para o plano de saúde.

O quadro abaixo sintetiza os artigos citados, anteriormente, e que serviram de base para a construção e raciocínio do presente estudo. Foram selecionadas as informações sobre qual dado e qual o tipo de aprendizado de máquina foram utilizados.

Quadro 1 – Artigos relacionados e informações sobre tipo de dados e de aprendizado de máquina utilizados

Autores	Utilizou aprendizado de máquina não supervisionado?	Utilizou dados de planos/seguros de saúde	Utilizou dados de saúde?
Araújo, Santana e Santos Neto (2016)	Não	Sim	Sim
Forkan, Khalil e Kumarage (2020)	Sim	Não	Sim
Meng et al. (2023)	Sim	Não	Sim
Santos, Dias e Chiavegatto Filho (2021)	Sim	Não	Sim
Seleme, Cubas e Carvalho (2023)	Não	Sim	Sim
Soleymani, Yaseri, Farzadfar, Mohammadpour, Sharifi e Kabir (2018)	Sim	Não	Sim
Xie, Schreier, Hoy, Liu, Neubauer, Chang, Redmond e Lovell (2016)	Não	Sim	Sim
Zhang, Li e McConnell (2021)	Não	Sim	Sim

Fonte: Autora (2023).

4 MATERIAIS E MÉTODOS

O capítulo apresenta o método utilizado nesta pesquisa para se modelar um processo de descoberta de conhecimento que atinja os objetivos deste estudo. A pesquisa, quanto a natureza, é considerada aplicada, pois, o produto proposto é de utilidade prática. A abordagem é qualitativa, visto o caráter exploratório e a visão subjetiva que estão presentes nas modalidades de análises de dados ditas descritivas e exploratórias utilizadas neste trabalho. Ele é descritivo porque seu objetivo principal é descrever os dados de entrada por meio da identificação de grupos com características semelhantes.

A análise de dados representa a combinação de tecnologia computadorizada, técnicas de ciência administrativa e estatística para solucionar problemas reais e o processo de descoberta de conhecimento em base de dados envolve uma sequência de etapas, como a limpeza de dados, na qual são tratados valores ruidosos e outliers (valores fora do limite) e a etapa de seleção, na qual somente os dados relevantes são filtrados para a pesquisa. Em seguida, os dados são transformados e consolidados de acordo com os propósitos da mineração. Realiza-se, então, a mineração de dados, no qual são aplicadas técnicas para descoberta de padrões nas bases de dados, por meio de algoritmos computacionais. Após a mineração de dados, inicia-se a etapa de análises de padrões que é precedida pela etapa de apresentação dos resultados e pela descoberta do conhecimento (FERNANDES et al., 2018).

O sucesso de um trabalho realizado utilizando técnicas de aprendizado de máquina se deve a três fatores segundo Doupe et al (2019). Primeiro, o problema a ser resolvido precisa estar bem delineado pela equipe de pesquisadores para se reduzir a chance de resultados falsos positivos; segundo, deve ser considerada a população que irá aplicar e analisar os resultados da pesquisa para que não ocorram interpretações equivocadas e terceiro, o banco de dados deve ser trabalhado e condizer com o problema a ser resolvido. Geralmente, o tamanho do banco não é importante e sim a qualidade dos dados selecionados.

4.1 MATERIAIS

A base de dados utilizada neste trabalho foi obtida, com o consentimento dos representantes contratantes do plano de saúde, exclusivamente, para fins de

pesquisa, sem qualquer identificação pessoal, a partir de uma base de teste de uma operadora do sistema de saúde suplementar contendo informações de procedimentos realizados em hospitais, laboratórios e consultórios relativas a beneficiários de um plano de saúde, disponibilizados na forma de um banco de dados em *Microsoft Excel*. Esses procedimentos obedecem à classificação exigida pela ANS (Agência Nacional de Saúde Suplementar) quanto ao padrão de troca de informação TISS (Troca de Informação em Saúde Suplementar).

A Troca de Informações na Saúde Suplementar - TISS, da agência nacional de saúde suplementar (ANS) - agência reguladora do setor de planos de saúde do Brasil, foi estabelecida como um padrão obrigatório para as trocas eletrônicas de dados de atenção à saúde dos beneficiários de planos, entre os agentes da Saúde Suplementar. O objetivo é padronizar as ações administrativas, subsidiar as ações de avaliação e acompanhamento econômico, financeiro e assistencial das operadoras de planos privados de assistência à saúde e compor o Registro Eletrônico de Saúde.

Exemplificando, a base de dados é composta de colunas que caracterizam o usuário do plano de saúde referente à sua data de nascimento e sexo. As outras colunas se referem aos procedimentos realizados por ele e onde foram realizados e são nomeadas como classificação SIP (sistema de informação de produtos) da ANS - que abrange as opções de consultas médicas, consultas médicas em pronto-socorro, demais despesas médico-hospitalares, terapias, exames simples, internações, demais exames, outros exames ambulatoriais e outras classificações como nome do prestador de serviço, descrição do procedimento/insumo, código especialidade procedimento - que abrange as opções de patologia clínica, tipos de cirurgias e áreas da medicina, por exemplo. Além desses itens consta a quantidade por módulo, o custo total do procedimento/insumo, o código especialidade principal - como análises clínicas, hospital, áreas da medicina - e a data de realização do procedimento/insumo.

Essa base é resultado dos vários serviços em saúde oferecidos pelas operadoras aos seus usuários sejam eles contratados de terceiros ou da própria rede médica. Quando um cliente realiza uma consulta, depois faz exames laboratoriais seguidos de exames de imagem, na maioria das vezes estes três procedimentos são realizados em prestadores diferentes. Todas as informações (referentes aos três procedimentos) são destinadas para a operadora com objetivo de realizar o pagamento aos prestadores (Fernandes e Raduenz (2020).

Os dados disponibilizados na base de procedimentos assistenciais em saúde

provenientes do sistema de saúde suplementar estão listados no Quadro 2. Esses dado se referem à base bruta (sem nenhum pré-processamento) disponibilizada pela operadora do sistema de saúde suplementar.

Quadro 2 - Estrutura da base de procedimentos assistenciais em saúde provenientes do sistema de saúde suplementar

Campo	Tipo	Exemplo	Possível atributo
Nome beneficiário	Textual	Fulano de tal	-
Grau parentesco	Textual	Titular	-
Data de nascimento	Data	01/01/1982	x
Sexo beneficiário	Categóricos	Masculino	x
Classificação SIP	Textual	Internações	x
Nome prestador	Textual	Hospital X	x
Descrição procedimento/insumo	Textual	Diária de quarto coletivo de 2 leitos com banheiro privativo	x
Quantidade por módulo	Numérico	1	x
Custo por módulo	Numérico	150,00	x
Código especialidade procedimento	Textual	Cardiologia	
Código especialidade principal	Textual	Hospital	x
Data de realização	Data	06/04/2022	x
Chave do documento	Numérico	00/00/000/TISS/000000/0	-
Código movimento	Numérico	0000000000	-
Período	Data	abril/2022	-

Fonte: Autora (2022).

4.2 METODOLOGIA

Em uma perspectiva de classificação dos dados, pertencentes a uma base de dados, podemos dividi-los em dados estruturados, não estruturados e semiestruturados. Os dados estruturados apresentam um formato pré- determinado e são armazenados em campos de dados que podem ser pesquisados, analisados e gerenciados. Eles podem ser divididos em dados categóricos (nominais e ordinais) e

numéricos (intervalares e racionais). Já os dados ditos não estruturados não podem ser facilmente categorizados. Embora tenham seu próprio sistema interno, eles não são considerados completamente consistentes e estruturados como um banco de dados e assim precisam sofrer algum processo de estruturação. Após esse processo, os algoritmos de aprendizado de máquina conseguirão agir sobre eles. São compostos por qualquer combinação de conteúdos textuais, de imagem, voz ou da Web. Os dados semiestruturados, por sua vez, possuem características de dois tipos descritos. Em essência, os dados estruturados são propícios para processamento por computador, ao passo que os dados não estruturados devem ser processados e compreendidos por humanos (técnicas de processamento de linguagem natural, por exemplo) (SHARDA et al., 2019, ABDELMAGID e QAHMASH, 2022 e SHARDA, DELEN E TURBAN, 2019).

Aos dados pré-processados foi aplicado algoritmo de agrupamento de aprendizado de máquina não supervisionado. O objetivo dessa técnica é obter uma melhor percepção do conjunto de dados e de como as diferentes variáveis se relacionam umas com as outras (BRUCE e BRUCE 2019).

O algoritmo utilizado neste estudo foi o K-means, considerado o algoritmo agregador mais empregado. Ele atribui cada ponto de dados a um agrupamento cujo centro ou centróide encontra-se mais próximo. O centro é calculado como a média de todos os pontos no agrupamento, ou seja, suas coordenadas são a média aritmética para cada dimensão separadamente por todos os pontos no agrupamento (SHARDA et al., 2019). O agrupamento é formado minimizando a soma das distâncias de cada atributo e o centro do seu grupo. O centro é o ponto equidistante dos indivíduos pertencentes ao conjunto (QUINTERO et al., 2022).

A quantidade de agrupamentos será observada de acordo com o valor da *silhouette*, do inglês, ou silhueta ou medida da qualidade da partição. Essa compara a distância média aos elementos no mesmo agrupamento com a distância média aos elementos em outros agrupamentos. Assim, ela sinaliza quão semelhante um objeto é ao seu próprio agrupamento (coesão) em comparação com outros agrupamentos (separação) (https://orangedatamining.com/widget-catalog/unsupervised/kmeans/).

O índice *silhouette* possui valores entre -1 e 1. Um valor próximo a 1 indica que o agrupamento foi bem formado e valores próximos a -1, indicam o oposto.

O valor mais alto indica que o objeto é compatível com seu próprio agrupamento e mal combinado com os agrupamentos vizinhos. Se a maioria dos objetos tiver um

valor alto, a configuração de agrupamento será apropriada. Se muitos pontos tiverem um valor baixo ou negativo, a configuração dos agrupamentos pode ter muitos ou poucos grupos. Logo, esse índice calcula duas medidas, uma em relação ao seu próprio agrupamento e outra em relação aos outros agrupamentos (dissimilaridade). Por fim, a estimativa do agrupamento global da base de dados se traduz em calcular a média desse escore nos pontos de interesse (QUINTERO et al., 2022; ŀKRLJ; KRALJ; LAVRAč, 2020)

Foi obtido um dendograma, como resultado do método hierárquico de agrupamento e foram analisados os agrupamentos formados.

O dendograma é construído a partir de uma abordagem aglomerativa ou divisiva. A aglomeração de dados inicia com a observação de cada variável na base de dados. Análises pareadas com resultados de menor dissimilaridade (há a necessidade de se definir uma medida de dissimilaridade) vão formar um agrupamento. Conglomerados similares são sequencialmente combinados em grupos maiores até formar um agrupamento que contém todos os dados. Já na técnica de formação de dendograma chama de divisiva, o algoritmo começa a analisar a base de dados como um todo e divide ela em hierarquias. A cada nível hierárquico um agrupamento é dividido para produzir dois novos grupos com grandes variações de características. Os agrupamentos são então sequencialmente divididos até que o número de grupos seja igual ao número de pontos de dados no conjunto de dados. O dendograma formado pode ser selecionado horizontalmente em diferentes níveis que formam coerentes agrupamentos (ECKHARDT et al., 2022).

4.3 FERRAMENTA

Para realizar essa dissertação foi utilizada a suíte de mineração de dados, chamada *Orange Data Mining* - Figuras 5.

O Orange Data Mining é um software de código aberto para mineração de dados baseado na linguagem *Python* em que o conhecimento é analisado e extraído de forma visual por meio de um conjunto de ferramentas inteligentes e de diversos algoritmos que auxiliam na compreensão dos dados. Ele possui uma interface gráfica fácil de usar e pode ser acessado pelo endereço eletrônico https://orangedatamining.com/ (ABDELMAGID e QAHMAASH, 2022).

Ele foi desenvolvido na Eslovênia no laboratório de bioinformática da

Faculdade de Ciências e Informação da Universidade de Liubliana em 1996 e oferece uma ampla gama de ferramentas estatístico-computacionais para análise de dados.

A escolha pelo Orange se deu pelo amplo material didático disponibilizado online pela empresa e por ser totalmente gratuito. Sua interface é fácil e parece ser uma ótima ferramenta para ser usada por profissionais que não fazem parte de carreiras da área da ciência da computação. Além disso, essa ferramenta foi utilizada em outros estudos de aprendizado de máquina de forma satisfatória por esta mestranda.

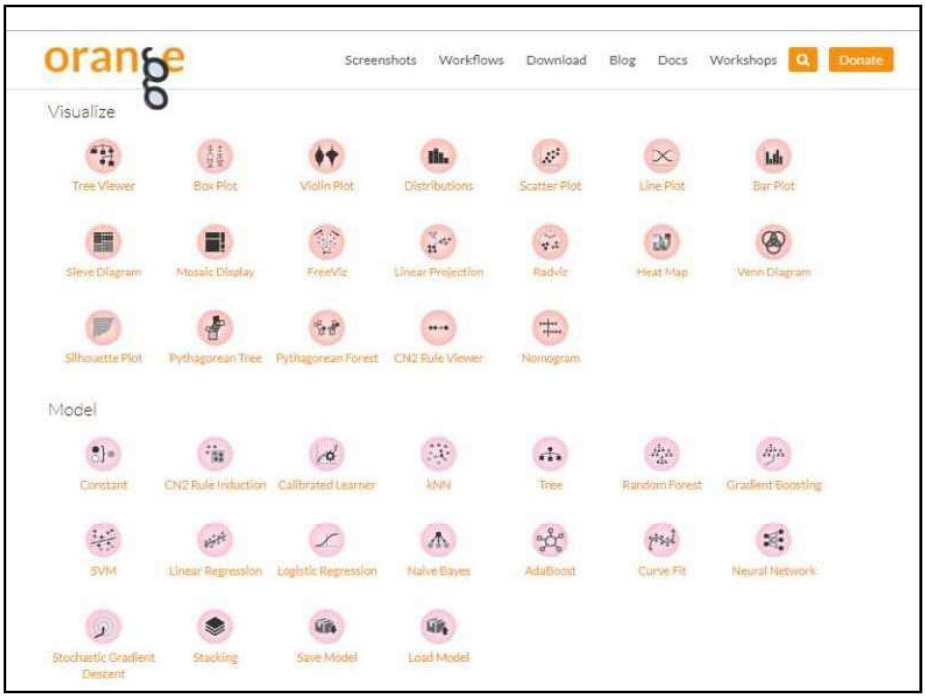
Dentre as ferramentas de descoberta de conhecimento e de mineração de dados que não exigem conhecimento avançado em alguma linguagem de programação como o *Python* e o *R* (amplamente utilizadas em análise de dados) e que possuem uma interface amigável, preferiu-se, frisa-se, o Orange Data Mining. Há opções de outros programas que buscam tornar a ciência de dados mais acessível a profissionais da área da saúde, por exemplo, sejam elas gratuitas ou pagas, como o Weka (https://www.cs.waikato.ac.nz/ml/weka/), o Rapidminer (https://rapidminer.com/), o Knime (https://www.knime.com/) e o Dataiku (https://www.dataiku.com/).

As figuras 2, 3 e 4 mostram algumas funcionalidades que a ferramenta disponibiliza por meio de Widgets. Estes são ícones que ficam disponíveis na tela de apresentação do programa e ficam organizados de acordo com a sua funcionalidade. Nos gráficos aqui citados, por exemplo, constam os widgets responsáveis pela obtenção dos arquivos de dados separados na parte de “Data”, assim como os ícones que correspondem às funcionalidades remetidas ao pré-processamento (*Transform*), à visualização (*Visualize*), ao aprendizado de máquina supervisionado (*Model*) e ao aprendizado de máquina não supervisionado (*Unsupervised*). Essas informações e imagens foram obtidas na página da internet da ferramenta e pode ser acessada por meio do endereço https://orangedatamining.com/widget-catalog/.

Figura 2 - Funcionalidades do Orange em relação aos dados e pré-processamento

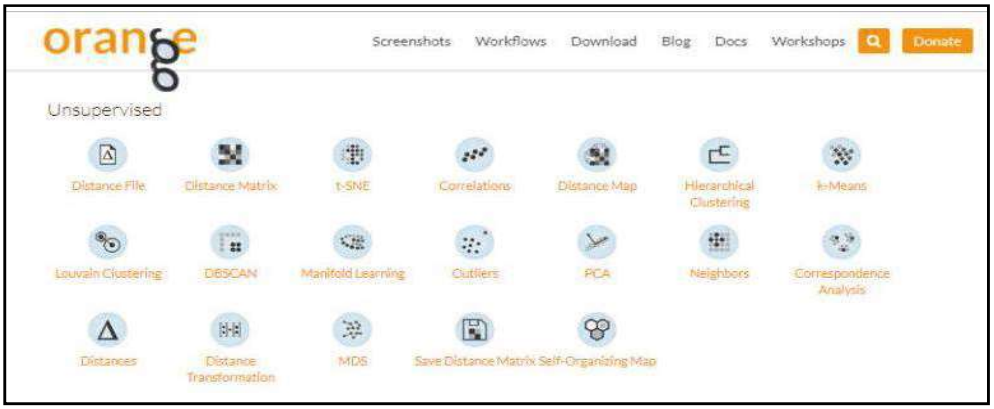
Fonte: Orange Data Mining (2023).

Figura 3 - Funcionalidades do Orange em relação às opções de visualização de resultados e em relação ao aprendizado de máquina supervisionado



Fonte: Orange Data Mining (2023).

Figura 4 - Funcionalidades do Orange em relação ao aprendizado de máquina não supervisionado

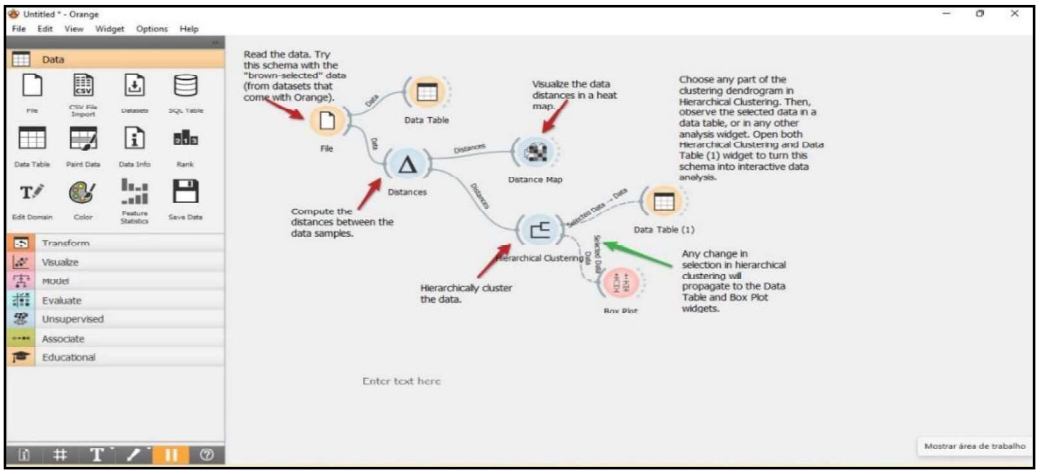


Fonte: Orange Data Mining (2023).

Os modelos de aprendizado de máquina e de análise de dados são formados arrastando-se os widgets para a área de trabalho do programa e fazendo-se a ligação entre eles de acordo com o objetivo do estudo.

A figura 5 mostra a tela do programa. As funcionalidades disponíveis estão à esquerda da imagem e está desenhado um modelo de agrupamento hierárquico. Esse modelo consta na base de apoio e de ensino das funcionalidades do programa disponibilizado nele, quando ocorre o download e na página do software no endereço https://orangedatamining.com/workflows/. Além dessas possibilidades de aprendizado, o Orange conta com um canal no *You Tube* https://www.youtube.com/@OrangeDataMining que disponibiliza muitos vídeos tutoriais sobre as suas funcionalidades.

Figura 5 - Modelo de análise no Orange Data Mining



Fonte: Orange Data Mining (2023).

Há widgets que oferecem a possibilidade de uso de bases de dados online, como *PubMed*, The Guardian, NY Times, Twitter, dentre outros, bastando que o usuário forneça, no caso do Twitter, a sua API. Caso o usuário possua sua própria base de dados, o Orange permite a importação de diversas extensões de arquivos, como .txt, .csv, .tsv, .xlsx, etc. O Orange oferece, ainda, um widget chamado Python Scripts, ampliando ainda mais as possibilidades de análise, pré-processamento e visualização dos dados (ALENCAR, 2023).

Os modelos de aprendizado de máquina não supervisionado disponibilizados no site e no canal do YouTube (https://www.youtube.com/@OrangeDataMining) foram exaustivamente estudados para se montar o modelo deste estudo.

4.4 AGRUPAMENTO

No caso de agrupamento realizado por meio do algoritmo k-means, no Orange Data Mining, os dados são analisados por suas similaridades ou dissimilaridades e agrupados por meio da medida de distância quadrática euclidiana. A distância quadrática euclidiana mede a distância entre duas observações (a, b) e todas as variáveis de cada observação (n), cuja fórmula é apresentada na equação abaixo – Equação 1 (SANTANA e PONTES, 2020).

$$\sqrt{\sum_{i=1}^{n}(A_i - B_i)^2}$$

Cada ponto, no caso em questão, cada código, deve pertencer a pelo menos uma partição. O algoritmo inicia selecionando aleatoriamente centros do conjunto de dados e, a cada interação, agrupa os pontos mais próximos ao centro por meio da menor distância quadrática euclidiana. Os centros são também atualizados pela média dos pontos de cada agrupamento. O algoritmo termina (converge) quando não há mudança significativa dos centros. (FERNANDES; CHIAVEGATTO FILHO, 2019).

O algoritmo divide os dados em “K” grupos por meio da minimização da soma das distâncias quadráticas de cada registro à média de seu grupo atribuído. Isso é chamado de soma dos quadrados dentro do grupo ou SS dentro do grupo. No agrupamento de registros com múltiplas variáveis o termo média do grupo não se refere a um único número, mas ao vetor das médias das variáveis. As medidas do K-médias não garantem que os grupos tenham o mesmo tamanho, mas encontra grupos

que sejam mais bem separados (BRUCE e BRUCE, 2019).

O funcionamento do algoritmo *k-means* pode ser resumido da seguinte forma, segundo Setiawan et al. (2022):

1) Escolha do número de agrupamentos;
2) Seleção aleatória de pontos iniciais de centros/centróides;
3) Atribuição de cada dado ao centro mais próximo de acordo com a distância euclidiana;
4) Recálculo do centro do agrupamento de acordo com os valores médios das distâncias (à medida em que as bases de dados vão aumentando);
5) Os passos 3 e 4 serão repetidos até que não haja mais mudança da distância dos pontos (representados como os dados) em relação ao centro de cada agrupamento.

Cada agrupamento será representado pelo seu centro, que é o ponto médio de todos os seus pontos e é utilizado para representar, descrever e resumir os dados de entrada (pontos) (SETIAWAN et al., 2022).

A base de dados utilizada foi anonimizada e organizada de maneira que o programa Orange Data Mining e a aplicação do algoritmo *K-means* pudessem ser executados da forma correta.

4.5 ESCOLHA DO ALGORITMO

Dentre as opções de algoritmos de agrupamento oferecidos pelo Orange Data Mining, o modelo que utiliza o *K-means* foi o mais apropriado, pois é o algoritmo que é utilizado em grandes bases de dados e com muitas variáveis como neste trabalho.

Embora haja outros algoritmos disponíveis na aba de funcionalidades relacionadas ao aprendizado de máquina não supervisionado como o DBSCAN e o *Louvain Clustering*, por exemplo, eles não são explicados em nenhum vídeo tutorial. Há inúmeras aulas sobre a utilização do algoritmo *K-means*, do T-SNE e do MDS, além do agrupamento hierárquico. Logo, a escolha sobre o modelo aqui proposto com a utilização do algoritmo *K-means* e do agrupamento hierárquico se justifica dessa maneira. Além disso, a mineração de dados utilizando o algoritmo *k-means* se mostrou satisfatória. Ainda, os agrupamentos formados apresentaram significado ao negócio e ao tema dos procedimentos que não se sentiu necessidade de se testar os outros algoritmos.

Segundo Bruce e Bruce (2019), o *k-means* ou K-médias foi o primeiro método de agrupamento desenvolvido, e ainda é muito usado, devendo sua popularidade à relativa simplicidade do algoritmo e sua habilidade de escalar grandes conjuntos de dados.

5 PROCESSO DE DESCOBERTA DE CONHECIMENTO EM BASE DE DADOS SOBRE PROCEDIMENTOS DE SAÚDE

Neste capítulo será abordado o processo de descoberta de conhecimento conforme Figura 1. Nele serão discorridas as atividades de análise da base de dados realizadas até a revelação de um modelo de aprendizado de máquina não supervisionado considerado mais apropriado e que responde de maneira satisfatória à questão de pesquisa.

Este capítulo foi dividido em três partes chamadas de pré-processamento, mineração e pós-processamento. A maior parte das informações se referem ao pré-processamento. Neste tópico constam as etapas do processo de DCBD acerca de como os dados foram selecionados, limpos e transformados demodo a apresentarem uma configuração ideal para a utilização pelo algoritmo. Na mineração consta a informação do algoritmo aplicado e a caracterização da base de dados e o resultado. O pós-processamento, neste capítulo, se resume a informar como os atributos de cada de agrupamento começaram a ser analisados.

A caracterização dos agrupamentos e a discussão acerca da aplicabilidade do modelo e do conhecimento descoberto constam no capítulo “Resultados e Discussão”.

5.1 PRÉ-PROCESSAMENTO

De forma diferente da base original, a tabela referência deste trabalho consta de colunas que retratam a idade dos usuários, o sexo e os códigos de sistema de informação de produtos (SIP) e de procedimentos principais dos procedimentos utilizados no plano de saúde. Essas duas últimas informações foram agregadas pois entende-se que a junção dessas informações caracterizou de forma apropriada e satisfatória os procedimentos de saúde.

Essa forma foi estabelecida pois os procedimentos na sua forma analítica, como a “Diária de quarto coletivo de 2 leitos com banheiro privativo”, resultam em muitas variáveis. Se o modelo utilizasse esses descritores - mais de três mil considerando várias dosagens de medicamentos, por exemplo, e seus valores, apresentou um tamanho de difícil análise e foi considerado inadequado.

Uma outra maneira testada de se caracterizar os procedimentos em saúde utilizados e aplicar a eles um modelo de aprendizado não supervisionado por meio da

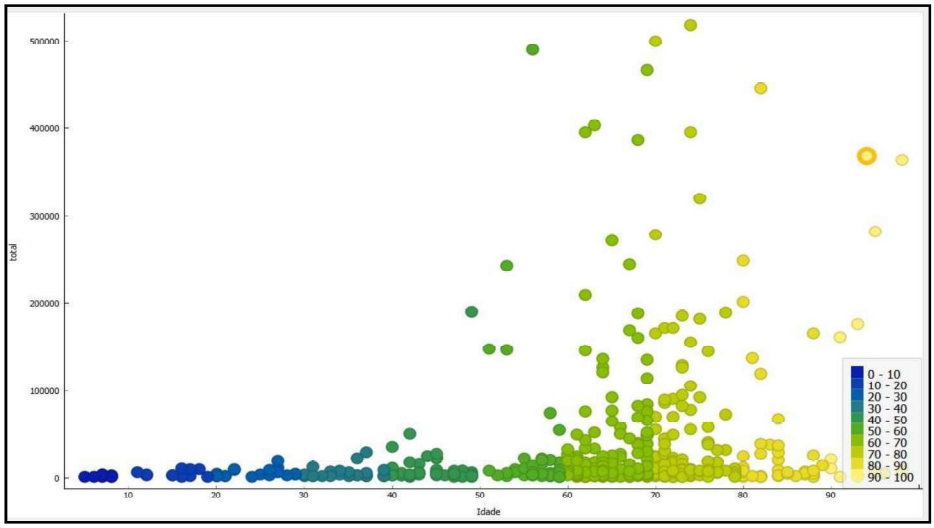
técnica de agrupamento foi utilizar os códigos de procedimento principal, que caracterizam a área médica do procedimento, como por exemplo, cardiologia, mas a quantidade de procedimentos sem essa classificação é muito grande e comprometeria a descoberta de conhecimento a respeito das características de utilização dessa carteira de beneficiários do plano de saúde.

Optou-se então por organizar a base de dados juntando a informação da SIP com o código de procedimento principal. Assim, os procedimentos ficaram caracterizados de acordo com os códigos SIP, terapias, consultas médicas em pronto socorro, e consultas médicas e com os códigos de procedimento principal como análises clínicas, hospital, cardiologia, serviços de imagem, clínica de cirurgia geral, fonoaudiologia, ortopedia e traumatologia, anestesiologia, por exemplo – anexos A e B. Apesar do nome dessas classificações conter “código”, eles são descritos em palavras como exemplificado no quadro 1.

A base de dados ficou composta de 3767 instâncias e de 267 atributos. Os usuários formam as linhas da tabela e os valores utilizados nos serviços de saúde foram alocados nas variáveis. Foi selecionado o período de doze meses, referentes ao ano de 2019 – primeiro ano completo de dados disponíveis - , sendo descartados os períodos posteriores a 2020 uma vez que, após o surgimento da pandemia da Sars-Cov-2, a utilização dos sistemas de saúde foi alterada substancialmente para atender aos casos ligados à essa doença. Optou-se por procedimentos deste ano por serem anteriores à pandemia da Sars-Cov-2. Como sabido, a utilização dos sistemas de saúde, após o surgimento do coronavírus, foi alterada substancialmente para atender aos casos ligados à essa doença. O comportamento de utilização dos usuários do plano de saúde está sendo retomado de forma gradual aos observados anteriormente a 2020.

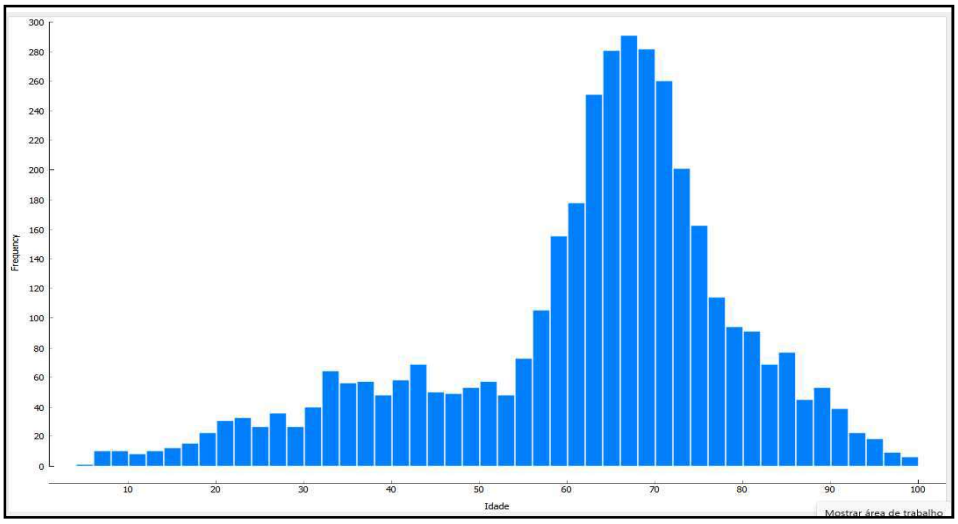
Abaixo, a figura 6, resume os valores utilizados anualmente pelos usuários e suas idades. Em análise conjunta com o gráfico de distribuição – figura 7 - percebe-se que a carteira é, predominantemente, composta de pessoas com mais de 50 anos. Um dos objetivos do presente trabalho é entender de que maneira esse público utiliza o plano de saúde e explorar possibilidades de ajudar essas pessoas a utilizarem o plano de saúde de forma correta (utilização de especialistas e de pronto-atendimentos de maneira assertiva e organizada) evitando desperdícios e reinternações desnecessárias e melhorando a navegação do cuidado em saúde dentro do sistema público e particular de saúde (YUILL; KUNZ, 2022).

Figura 6 - Custo X Idade



Fonte: Autora (2023).

Figura 7 - Distribuição de usuários por idade

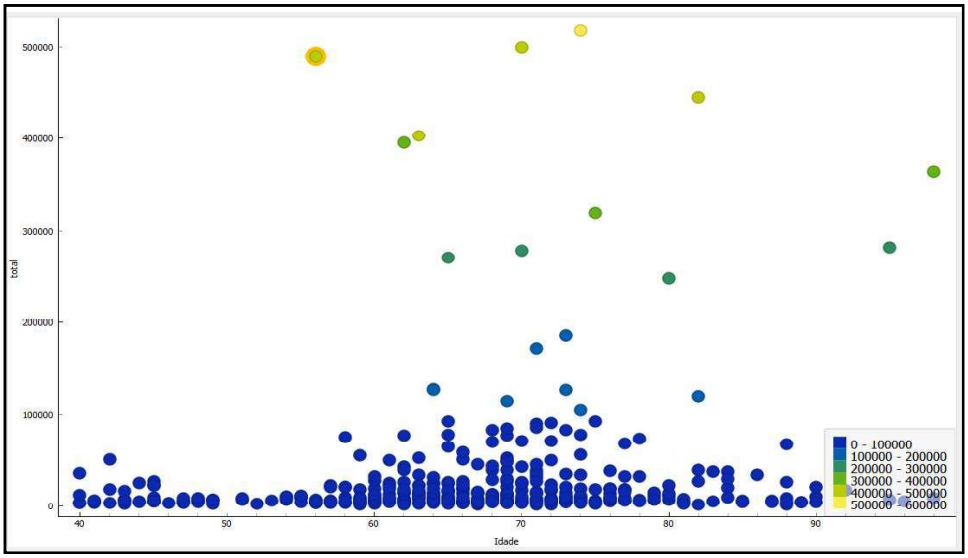


Fonte: Autora (2023).

Por conseguinte, os usuários com menos de quarenta anos (figura 8) foram excluídos da base, assim como os usuários considerados “outliers”. As pessoas com gastos excessivos no sistema de saúde são facilmente localizadas em uma base de dados com ajuda de softwares como o *Microsoft Excel* e não necessitam, de acordo com os objetivos deste trabalho, de uma análise mais apurada com o uso de modelos de aprendizado de máquina.

[1] Os gráficos demonstrados por meio das Figuras foram elaborados utilizando o programa Orange Data Mining.

Figura 8 - Idade e total de usuários com mais de 40 anos



Fonte: Autora (2023).

Após a análise dos valores mínimos e máximos utilizados assim como a frequência em que foram utilizadas as variáveis e sua interferência na visualização do dendograma das variáveis e do resultado dos segmentos encontrados na aplicação do algoritmo K-means, optou-se por excluir as variáveis com pouca frequência (cerca de 5) e valores muito baixos (menos de

100 reais) e os usuários que tiveram um custo menor de R$1.000,00, independentemente da idade. Ainda, considerando a figura 8 foram excluídos os usuários que tiveram gastos em saúde com valores maiores de duzentos mil reais, estes considerados como outliers. Assim, restaram na base de dados usuários que representam a maior fatia de faixa de custo de utilização com a maior quantidade de procedimentos realizados – figuras 14, 17 e 18. Isso foi realizado como uma forma de deixar a base de dados mais homogênea para que o algoritmo possa atuar de forma a demonstrar resultados melhores. Entende-se que uma base menos homogênea pode contribuir para que o algoritmo ofereça resultados menos satisfatórios e focados em variáveis responsáveis pelas discrepâncias da base. O total de usuários excluídos em virtude dessas particularidades totalizou 734 – quadro 2.

O que se quer com esse trabalho é conseguir descobrir conhecimento para a tomada de decisão em gestão de saúde a respeito da grande quantidade de usuários que faz uso do sistema de saúde suplementar de forma parecida. Pacientes que possuem um custo e idade facilmente identificáveis, como os do gráfico 1, na faixa etária de 10 a 20 anos, por exemplo, não precisam ser caracterizados por meio de

algoritmos de aprendizado de máquina quando o objetivo do estudo é descobrir padrões de uso em uma base de dados cuja maioria dos usuários é idosa.

Como etapas de pré-processamento da base de dados podemos citar:

1) Seleção de procedimentos referentes ao ano de 2019: foi considerado um ano inteiro de procedimentos, ou seja, 12 meses, por se entender que é o suficiente para se começar a entender características de utilização de procedimentos em saúde em um grupo de pessoas majoritariamente idosa;

2) Anonimização dos dados: no Brasil, a privacidade dos dados está regulada pela lei de nº 13.709, de 14 de agosto de 2018 - Lei Geral de Proteção de Dados (LGPD). A autoridade nacional de proteção de dados é o órgão responsável por regular a aplicação da lei e das sanções impostas por ela.

Esta lei protege com mais rigor os dados pessoais sensíveis, os quais se referem a "origem racial ou étnica, convicção religiosa, opinião política, filiação a sindicato ou a organização de caráter religioso, filosófico ou político, dado referente à saúde ou à vida sexual, dado genético ou biométrico, quando vinculado a uma pessoa natural". Trata-se de informações que interessam tão somente às pessoas das relações privadas ou íntimas do indivíduo ou, em alguns casos, exclusivamente ao próprio indivíduo. Como são informações próprias da intimidade, são consideradas sensíveis, motivo pelo qual merecem proteção jurídica. O tratamento de dados pessoais sensíveis que viole as regras de proteção à privacidade pode provocar danos materiais e morais ao seu titular. Em matéria de saúde, por exemplo, pelo conhecimento de certas doenças é possível, por exemplo, inferir sobre a sexualidade de um paciente. Essa inferência, acertada ou não, pode afetar suas relações familiares, seus relacionamentos íntimos e até mesmo sua vida profissional, especialmente no caso de pessoas públicas (ANS, 2019).

A base de dados utilizada neste trabalho contém dados de saúde. Esses dados são considerados sensíveis pela Lei Geral de Proteção de Dados e merecem ser tratados de acordo com ela. Ou seja, os dados, que possivelmente possam identificar usuários, devem ser anonimizados. Além disso, a lei estabelece que todas as pessoas que manipulam esses dados devem seguir regras de segurança da informação para não haja vazamentos e danos para os usuários. O percurso do dado, quem está fazendo uso dele e por que deve ser passível de monitorização e rastreamento de forma que os responsáveis por possíveis infrações sejam penalizados.

Seleção das colunas código, sexo, idade, valor total do procedimento, código

SIP e código de procedimento principal: essa seleção de colunas se fez necessária pois a quantidade de procedimentos encontrados na base de dados atingiu um número muito grande (3150). Além disso, dentre os procedimentos constam materiais como agulhas, gaze, esparadrapo, entre outros que somam grandes quantidades e que não apresentam potencial de caracterizar grupos de utilizadores de procedimentos em saúde. As colunas sexo e idade são importantes para caracterizar pessoas e alguns procedimentos em saúde e doenças estão relacionados à idade, como doenças neurológicas como o Alzheimer e ao sexo como câncer de ovário que atinge mulheres e câncer de próstata, que atinge os homens;

3) Junção das colunas código SIP e código de procedimento principal: essa união se mostrou interessante para se caracterizar um procedimento. Por exemplo, um procedimento como o cateterismo, estará configurado na base de dados como Internações (SIP) e Cardiologia (código de procedimento principal). Assim, conseguiu-se reduzir a quantidade de variáveis sem perder a essência do procedimento e a descoberta de conhecimento sobre as áreas médicas de utilização do plano de saúde;

4) Transformação da tabela original:

- cada descrição formada na junção dos códigos SIP e dos códigos de procedimento principal formou um atributo;
- cada código formou uma linha;
- a planilha ficou organizada como as linhas sendo os códigos dos usuários e as colunas representando as variáveis que descrevem os procedimentos em saúde realizados pelos usuários acrescentadas das colunas que representam a idade, o sexo, o valor dos procedimentos e o valor total dos custos.

A planilha original caracteriza um procedimento em uma coluna pelo seu nome. O nome da coluna é “Descrição do procedimento”. No modelo aqui proposto cada um desses procedimentos foi agrupado em seu código de procedimento principal e na sua SIP. Reitera-se que essa classificação está na base de dados original conforme nos mostra o quadro 2. O que foi feito foi utilizar outras classificações dos procedimentos como forma de agrupá-los sem perder sua essência.

5) Exclusão dos códigos com informação de idade com número menor de 40: figura 7 nos mostra a distribuição dos códigos por idade. Como o público desse grupo é majoritariamente idoso, ou seja, possuem mais de 60, são as pessoas que mais gastam – figura 6, um grupo homogêneo e com mais condições de se encontrar

melhores resultados com a aplicação de um algoritmo pode ser formado assim. Ainda, observa-se claramente uma curva Normal ao se retirar esses códigos o que nos favorece com uma distribuição de pessoas conforme a idade de uma maneira propícia a se conseguir melhores resultados analíticos – figura 13;

O estatuto do idoso, cuja lei é a de número 10.471 do ano de 2003, em seu artigo primeiro, define idoso como as pessoas com idade igual ou superior a 60 (sessenta) anos.

6) Exclusão dos códigos com informação de custo total menor de R$1.000,00: a figura 6 nos mostra um panorama dos códigos que possuem custo total de até mil reais. O motivo principal de exclusão desses códigos é a baixa utilização anual do plano. Esses usuários não fazem parte do escopo deste estudo pois não onera o plano de saúde de forma significativa e provavelmente utilizam o plano para procedimentos e exames de baixa complexidade e para consultas devido a problemas pontuais. Se formos considerar que uma consulta com profissional médico custa em torno de R$100,00 - cem reais - (pesquisa realizada na própria base de dados por descrição de procedimento) o valor de mil reais não seria considerado exagerado.

7) Exclusão dos códigos com informação de custo total maior de R$200.000,00: analisando-se a figura 6 vislumbram-se claramente códigos com valores de custo total em planos de saúde maiores que duzentos mil reais. Esses códigos são facilmente identificáveis e sua conta estudada sem a necessidade de utilização de um modelo de aprendizado de máquina para ajudar. Além disso, entende-se que esses códigos podem contribuir para algum tipo de viés quando da aplicação do algoritmo de agrupamento. A formação de um agrupamento, somente com esses usuários, não seria uma descoberta interessante e original pois esses valores podem ser facilmente identificados em ferramentas da *Microsoft* como o *Excel* ou o *Power BI*.

8) Exclusão de variáveis cujo valor total (considerando todos os códigos) somou menos de R$100,00;

9) Exclusão de variáveis cuja incidência tenha sido inferior a 10 códigos.

A exclusão das variáveis cujo valor total tenha sido menos de cem reais e das variáveis cuja incidência tenha sido inferior a 10 códigos foi definida a partir da base de dados transformada e formatada no software *Microsoft Excel*. Essa base foi testada no modelo construído na ferramenta Orange Data Mining.

A análise da distribuição das variáveis tanto na formação do dendograma quanto na formação dos agrupamentos não apresentou resultados satisfatórios e

optou-se por retirar variáveis com pouca incidência e valores baixos para não comprometer a homogeneidade do modelo e não formar vieses. A finalidade do modelo aqui proposto é descobrir conhecimento a partir de códigos que são responsáveis por caracterizar o perfil de utilização do plano de saúde pelo custo e pelas várias áreas médicas referenciadas.

Foram realizados filtros, no *Microsoft Excel*, para se encontrar quais seriam os valores totais de cada atributo, analisando-se a carteira toda, e qual a frequência dos valores, ou seja, quantos códigos fizeram uso dos procedimentos qualificados no atributo referência. Assim, optou-se por retirar as variáveis com essas especificações por apresentarem baixa frequência e valores.

Quadro 3 - Quantitativo de instâncias e atributos

	Base original	**Base pré-processada**
Instâncias	3767	267
Atributos	3033	165

Fonte: Autora (2022).

Quadro 4 - Etapas do Pré-processamento

ETAPAS PRÉ-PROCESSAMENTO
- Seleção de meses e ano de procedimentos;
- Anonimização dos dados;
- Seleção das colunas código, idade, sexo, código SIP e código procedimento principal;
- Transformação da tabela original;
- Exclusão dos códigos com informação de idade menor de 40;
- Exclusão dos códigos com informação de custo total menor de R$1.000,00;
- Exclusão dos códigos com informação de custo total maior de R$200.000,00;
- Exclusão das variáveis cujo valor total (considerando todos os códigos) totalizou menos de R$100,00;
- Exclusão de variáveis cuja incidência tenha sido inferior a 10 códigos.

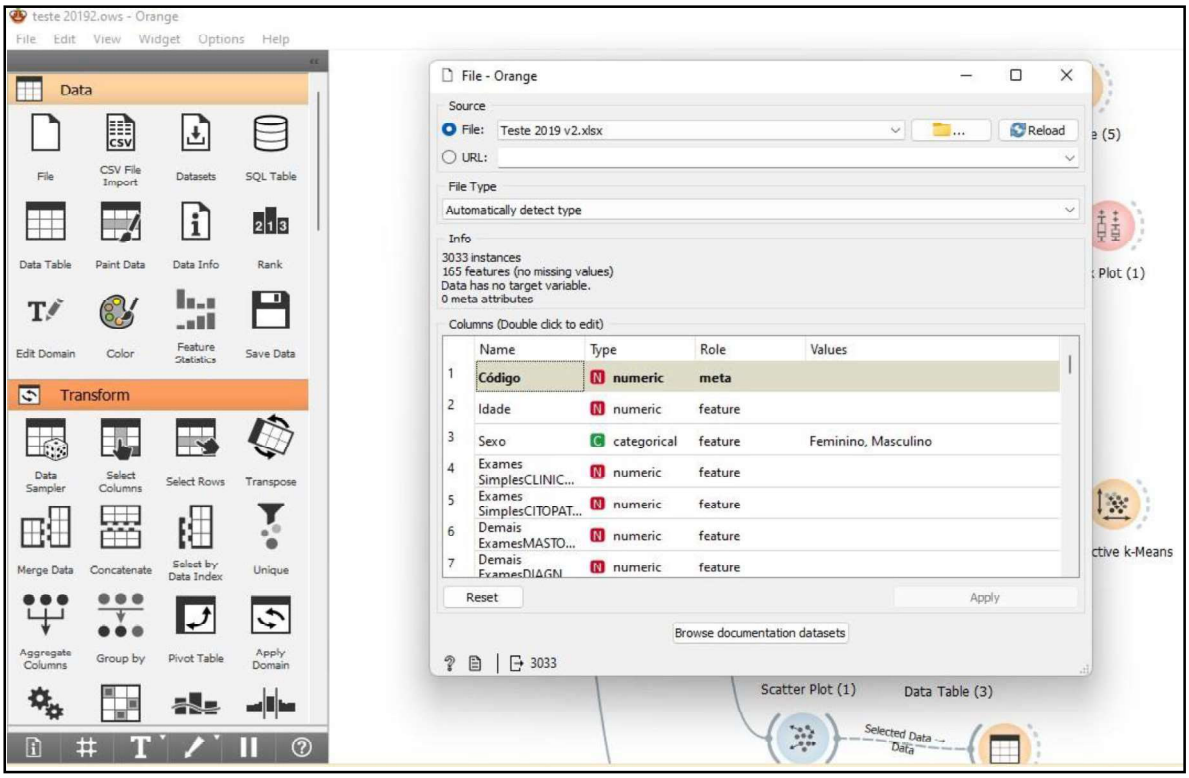
Fonte: Autora (2022).

5.2 MINERAÇÃO DE DADOS

O algoritmo de agrupamento *k-means* foi aplicado à base de dados cujas informações constam na figura 9. Repara-se que não há valores faltantes na base de

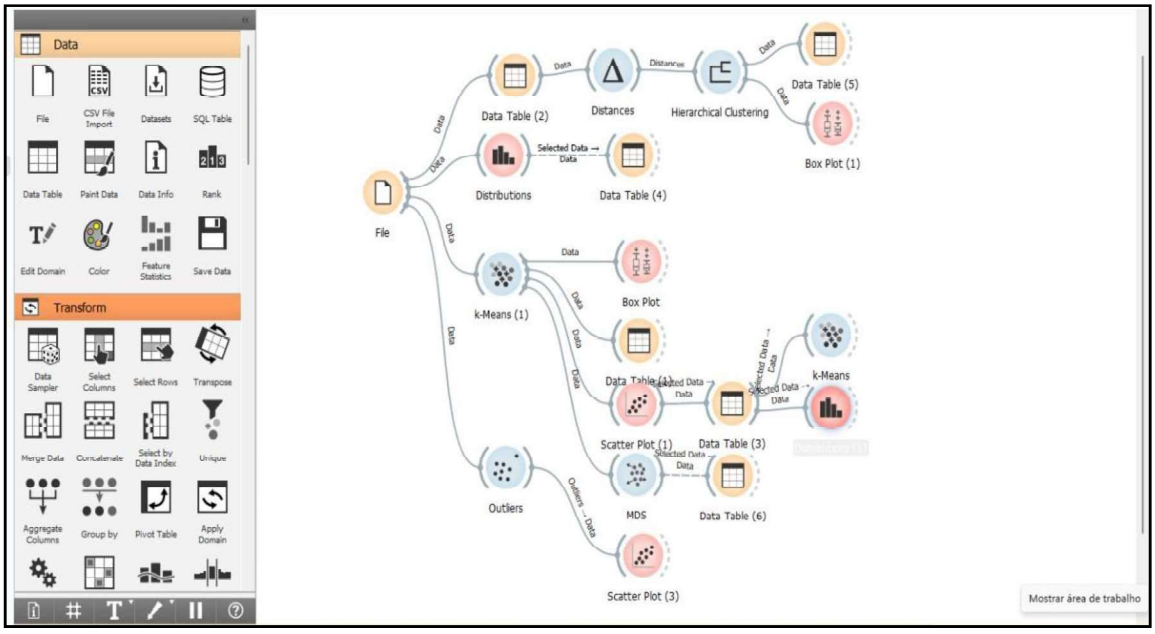
dados e as variáveis em sua maioria são numéricas pois retratam o seu custo. A variável sexo que foi classificada como categórica. E o modelo utilizado junto à ferramenta de mineração de dados consta na figura 10.

Figura 9 - Informações sobre a base de dados



Fonte: Orange Data Mining (2023).

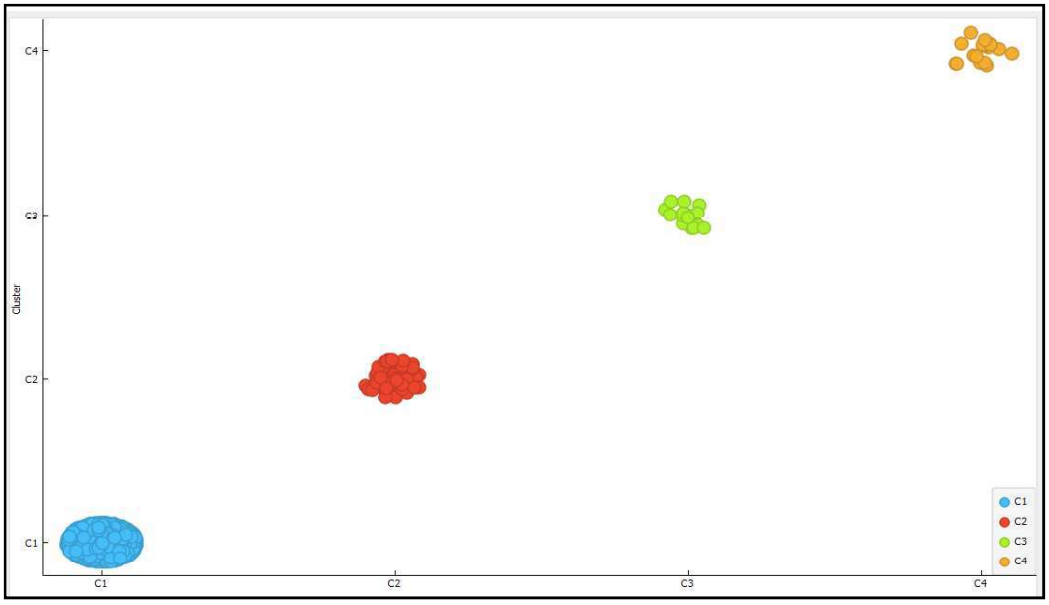
Figura 10 - Modelo de mineração de dados



Fonte: Orange Data Mining (2023).

O modelo descritivo (figura 11) que apresentou melhores resultados caracterizou os códigos em quatro grupos – figura 12. A quantidade de usuários em cada agrupamento variou bastante sendo que no agrupamento 1 tem-se 2929 integrantes, o 2, 71, o 3, 15 e o 4 tem 18. O agrupamento 1 é o mais numeroso dos 4. Possui 2929 integrantes, cerca de 78% de todos os usuários. Como uma maneira de se analisar os participantes desse grupo e de se testar se havia a possibilidade de alguma subdivisão dele, foi aplicado o algoritmo *K- means* sobre ele. Como resultado obteve-se dois grupos. Um numeroso e outro com poucos usuários. Não foi vislumbrada nenhuma particularidade que justificasse a aplicação e a demonstração desses subgrupos. Portanto, o modelo com quatro agrupamentos, sendo um deles bem numeroso em relação aos demais, se mostrou interessante e aplicável.

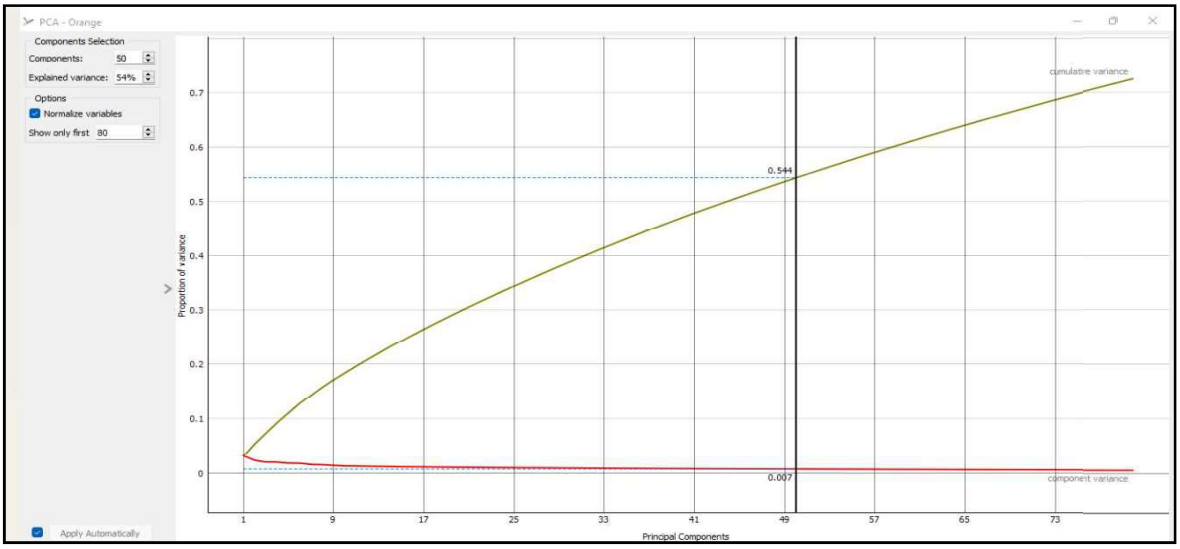
Figura 11 - Agrupamentos



Fonte: Autora (2023).

Como forma de se constatar quais são as principais variáveis responsáveis pelo resultado dos agrupamentos foi aplicado o algoritmo de análise de componentes principais. O resultado consta na figura 9. Cinquenta componentes explicam 50% das variáveis. A quantidade de componentes encontrada foi bem grande e a quantidade de agrupamentos formados aplicando-se esse algoritmo resultou em quatro. Mesmo número do modelo sem utilizar a redução de dimensionalidade.

Figura 12 - Análise de componentes principais



Fonte: Autora (2023).

5.3 PÓS-PROCESSAMENTO

Neste item são abordadas as etapas dos padrões extraídos e de que forma eles poderão ser aplicados de forma a melhorar a saúde da população analisada. Esses tópicos serão aprofundados no capítulo seguinte denominado “Resultados e Discussão”.

A base de dados foi analisada de forma global utilizando-se as ferramentas de detecção de outliers e de distribuição das variáveis, por exemplo, para que se tenha uma melhor compreensão das informações encontradas e se cumpra os objetivos de um modelo de processo de descobrimento. Ou seja, foram analisadas todas as variáveis em relação aos seus resultados por agrupamentos e não somente as variáveis principais (figuras 31, 38, 47 e 51).

Este processo de descoberta de conhecimento em base de dados é descrito em etapas sequenciais, ou seja, inicia-se com a seleção dos dados e segue para o pré-processamento, transformação dos dados, mineração dos dados e aplicação do conhecimento (figura 1), mas ele é interativo e iterativo. Até a descoberta do modelo aqui proposto os dados foram testados de diversas maneiras, como exemplo dessas tentativas podemos citar as etapas de escolher quais seriam as faixas etárias escolhidas, como um modelo se comportaria se deixássemos todas as variáveis considerando sua frequência e valor total. Ainda, se considerássemos todos os códigos sem observar o mínimo de procedimento utilizado no período proposto e os seus

valores totais.

Quadro 5 - Etapas do DCBD e atividades

ETAPA	ATIVIDADES
Seleção dos dados	Escolha de 12 meses de utilização do ano de 2019
Limpeza dos dados	Exclusão de colunas da base original. Exemplo, nome do prestador de serviço
Transformação dos dados	As variáveis escolhidas são apresentadas em forma de linhas na base original. Elas precisaram ser transformadas em colunas
Mineração dos dados	Aplicação do algoritmo K-means
Internalização	Analisar se os agrupamentos formados fazem sentido de acordo com a pergunta de pesquisa e se podem ser melhorados

Fonte: Autora (2023).

Assim, no quadro de número 4 consta as etapas do processo de descobrimento em base de dados e alguns exemplos de atividades realizadas neste trabalho até o atingimento do melhor modelo.

6 RESULTADOS E DISCUSSÃO

Nesta seção serão abordados os resultados encontrados na aplicação do algoritmo K-means na base de dados explicada na seção anterior. Além da quantificação dos agrupamentos e suas caracterizações será apresentado um dendograma que será analisado de forma conjunta aos resultados do algoritmo e os resultados das análises realizadas com o auxílio das funções de outliers e distribuição de frequência das variáveis.

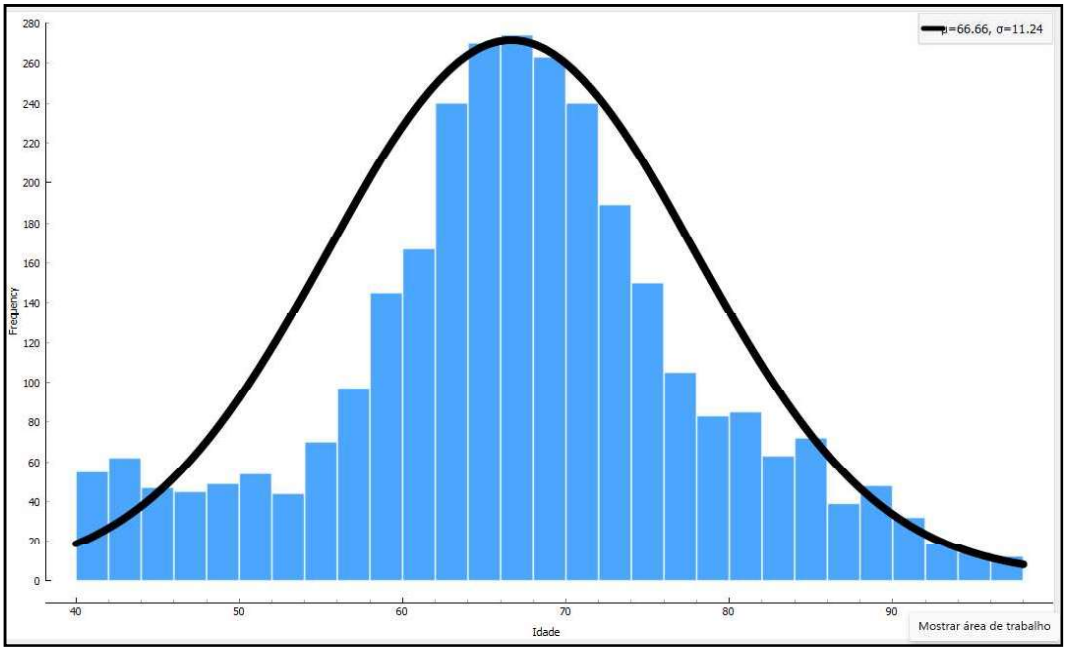
Uma ampla quantidade de gráficos foi analisada conjunta e separadamente entre os agrupamentos para que o melhor conhecimento acerca da base de dados e sua segmentação fosse descoberto.

6.1 ANÁLISE CONJUNTA DOS DADOS E AGRUPAMENTOS

A base de dados utilizada, depois da etapa de pré-processamento, ficou composta por 3033 códigos e 165 variáveis - figura 10. A distribuição das idades ficou resumida na figura 13. 1287 usuários possuem de 62 a 72 anos o que corresponde a 42,43% de toda a população estudada. Analisando-se em conjunto as figuras 13 e 14 podemos verificar que eles se assemelham e que os custos maiores e mais variados se encontram nesta faixa etária que abarca um período de 10 anos.

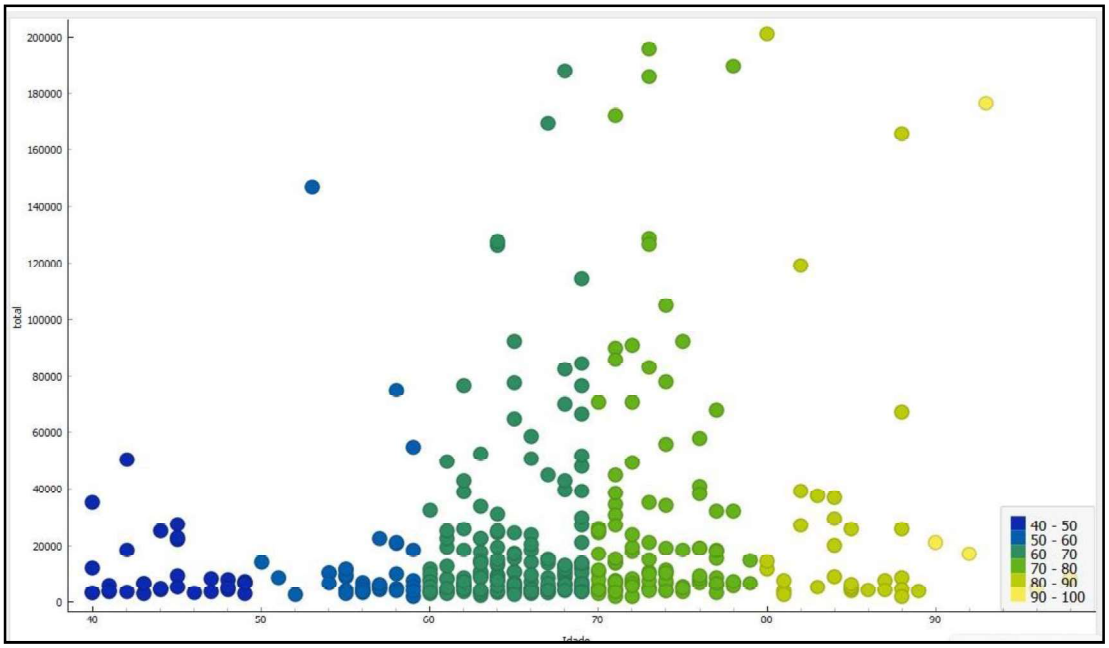
A figura 16 nos mostra a divisão dos usuários por faixa etária. Percebe-se que a quantidade de mulheres atinge um número próximo de 2 mil e o de homens, um pouco acima de 1 mil, o que significa que a quantidade de mulheres é quase o dobro da quantidade de homens. Procedimentos relacionados à saúde feminina como os relacionados à ginecologia e mastologia podem sofrer mais impacto dentro de uma análise conjunta de variáveis que procedimentos ligados à saúde masculina como os ligados à urologia, por exemplo.

Figura 13 - Representação da quantidade de usuários pela idade



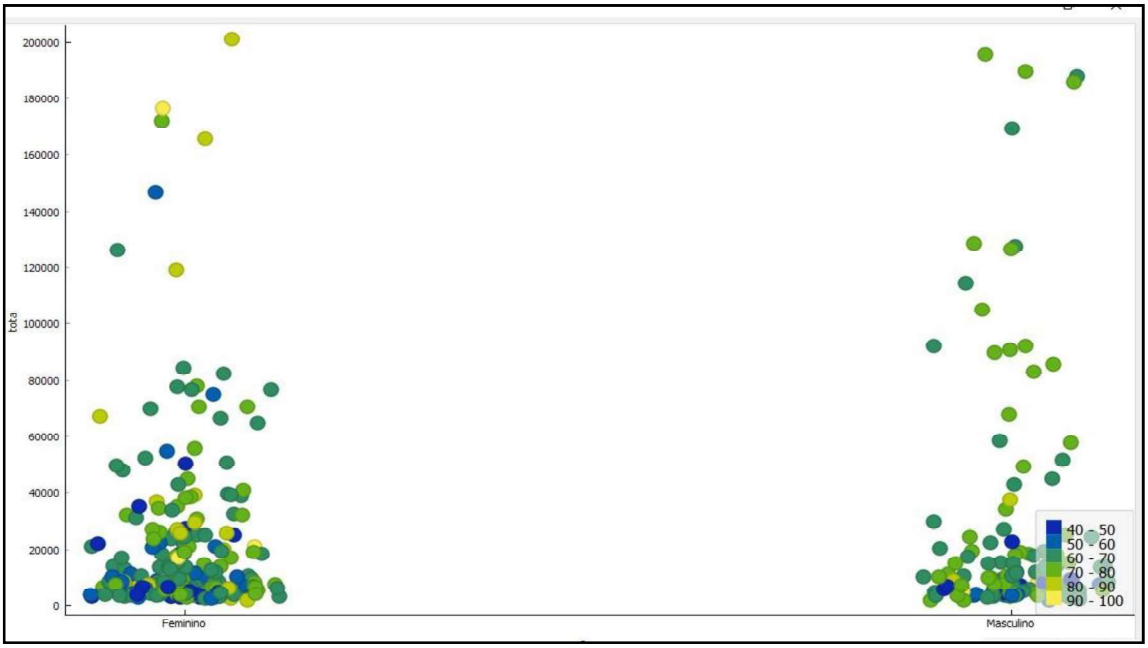
Fonte: Autora (2023).

Figura 14 - Representação da idade e custos



Fonte: Autora (2023).

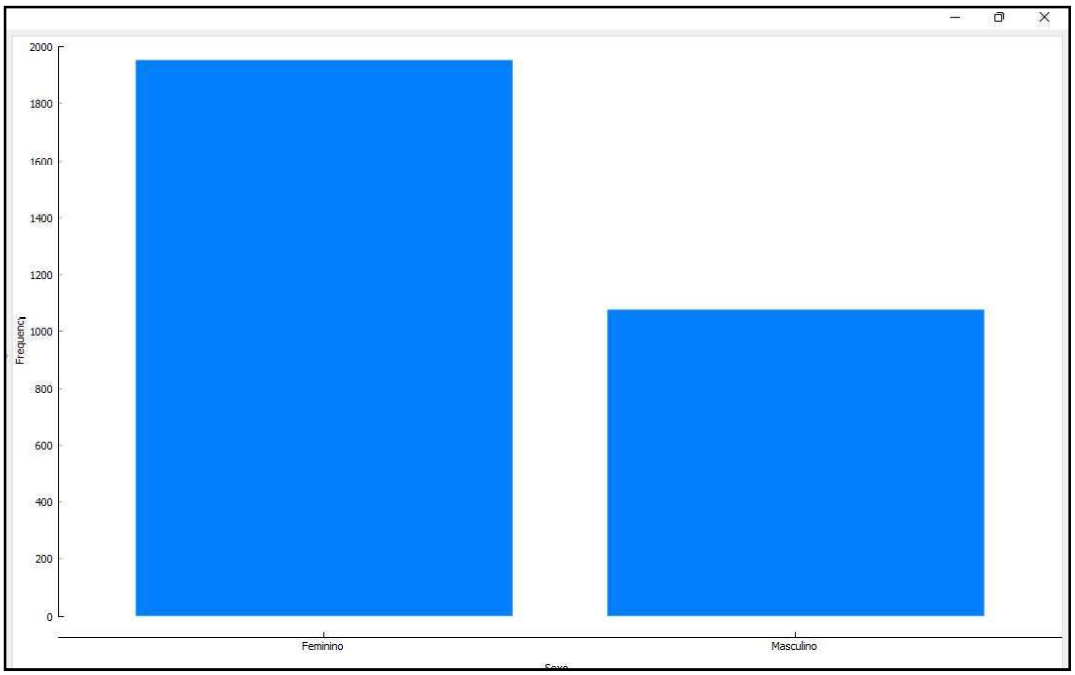
Figura 15 - Representação dos custos por sexo e idade



Fonte: Autora (2023).

Analisando-se a figura 15, constata-se que os usuários do sexo feminino que gastaram mais de R$100.000,00 possuem em sua maioria mais de 70 anos enquanto os usuários do sexo masculino com essa mesma característica monetária possuem idade menor.

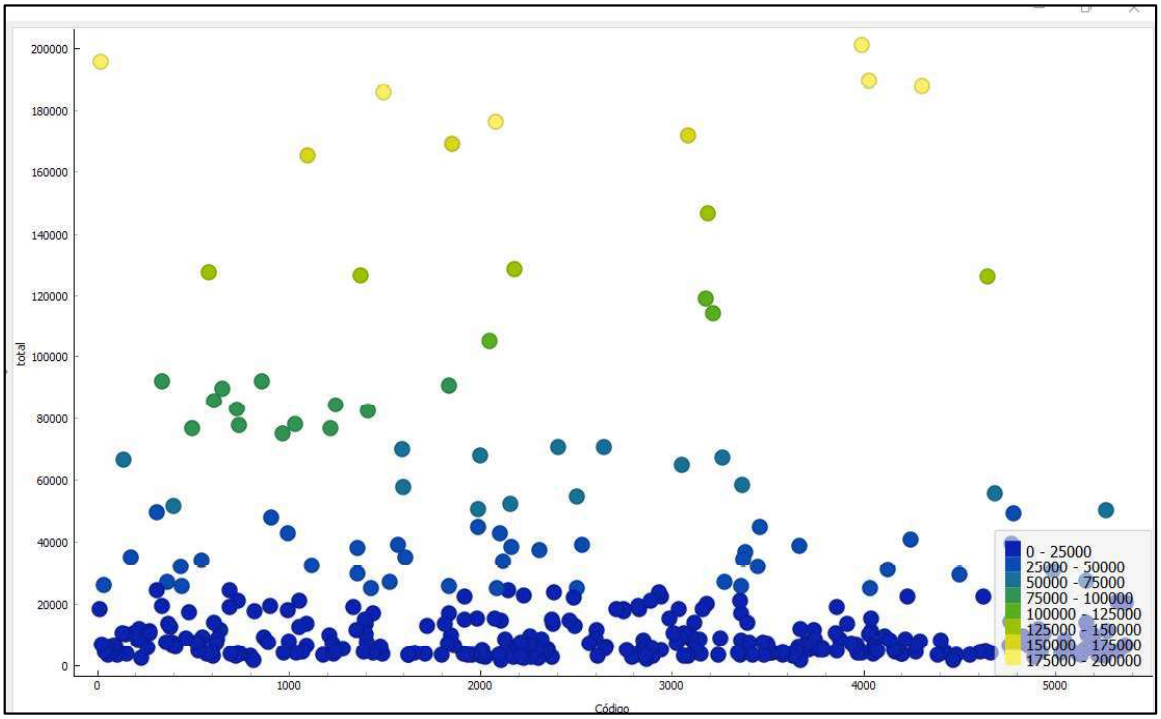
Figura 16 - Representação da quantidade de usuários por sexo



Fonte: Autora (2023).

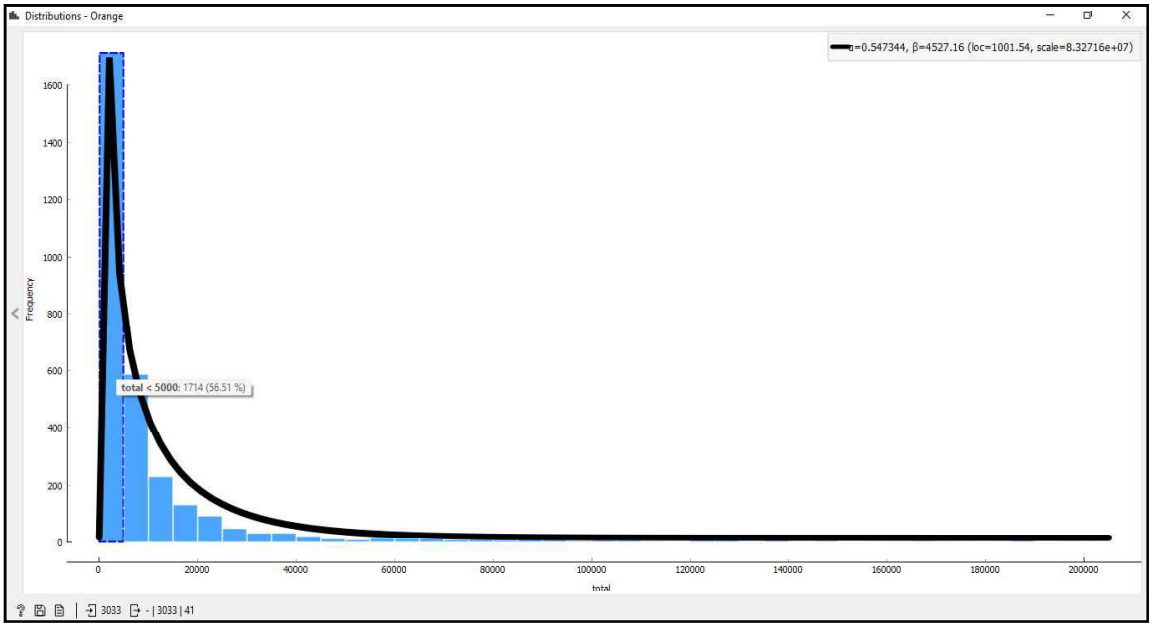
A figura 18 nos mostra a divisão dos usuários do plano de saúde por custo despendido em um período de um ano. Mais da metade dos usuários, ou seja, 75,87% (2301) gastaram até R$10.000,00 (dez mil reais). A frequência de pessoas que gastaram valores mais altos diminuiu consideravelmente a partir da faixa que compreende valores entre R$40.000,00 e R$45.000,00. Experimentalmente, tentou-se excluir os códigos das pessoas que utilizaram o plano em valores maiores de R$40.000,00, mas concluiu-se que os códigos restantes poderiam não expressar de maneira correta o comportamento de gastos da carteira e quais foram os procedimentos utilizados. Os verdadeiros valores considerados outliers foram, anteriormente, retirados da base de dados e gastos de até R$200.000,00 não seriam considerados tão expressivos quando se trata de uma carteira de idosos com alta probabilidade de precisarem realizar alguma cirurgia e sujeitos a tratamentos dispendiosos. As figuras 14 e 17 esclarecem que os maiores valores não comprometem a análise da carteira e podem ser considerados pertencentes ao mesmo grupo de utilizadores.

Figura 17 - Custo x códigos da base de dados utilizada



Fonte: Autora (2023).

Figura 18 - Frequência de usuário por custo



Fonte: Autora (2023).

À base de dados pré-processada foi aplicado o algoritmo de aprendizado não supervisionado *K-means* – figura 11. A quantidade de grupos formada baseou-se na métrica *silhouette* conforme demonstra a figura 19. Quatro agrupamentos foram formados e constam demonstrados na figura 12.

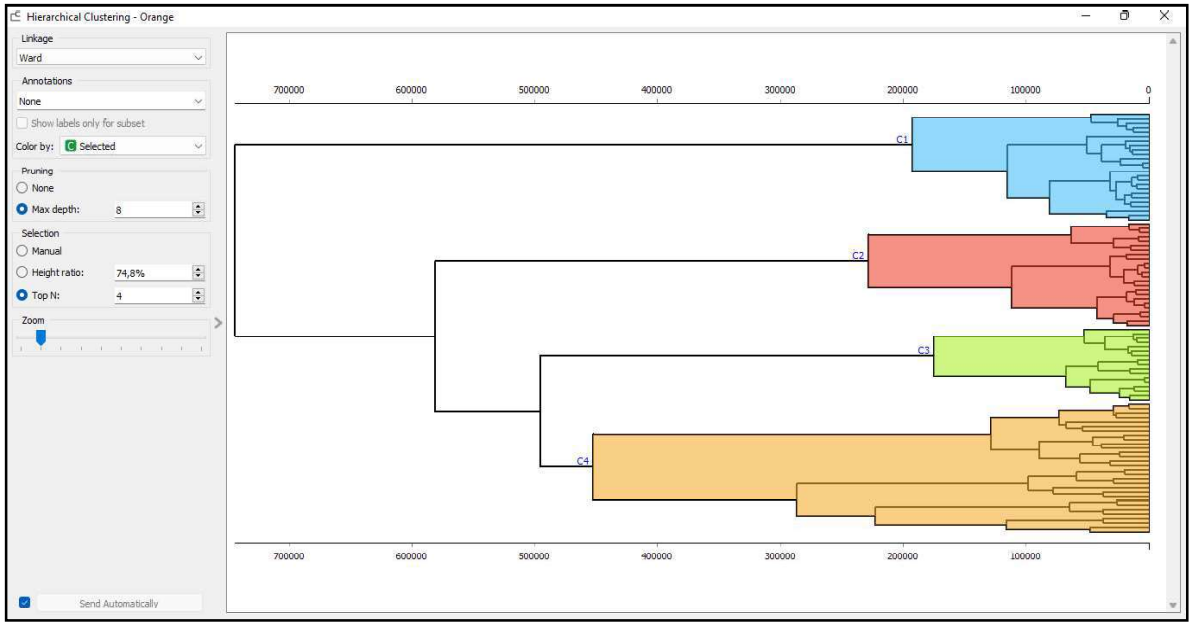
Figura 19 - Avaliação Silhouette

Fonte: Autora (2023).

O dendograma encontrado consta na figura 20. Nele podemos observar os agrupamentos separados por cores, o que nos ajuda com a visualização do tamanho

deles e o grau de similaridade de utilização entre os usuários. Verticalmente podemos verificar como as variáveis estão separadas por agrupamento.

Figura 20 - Dendograma com a representação dos agrupamentos



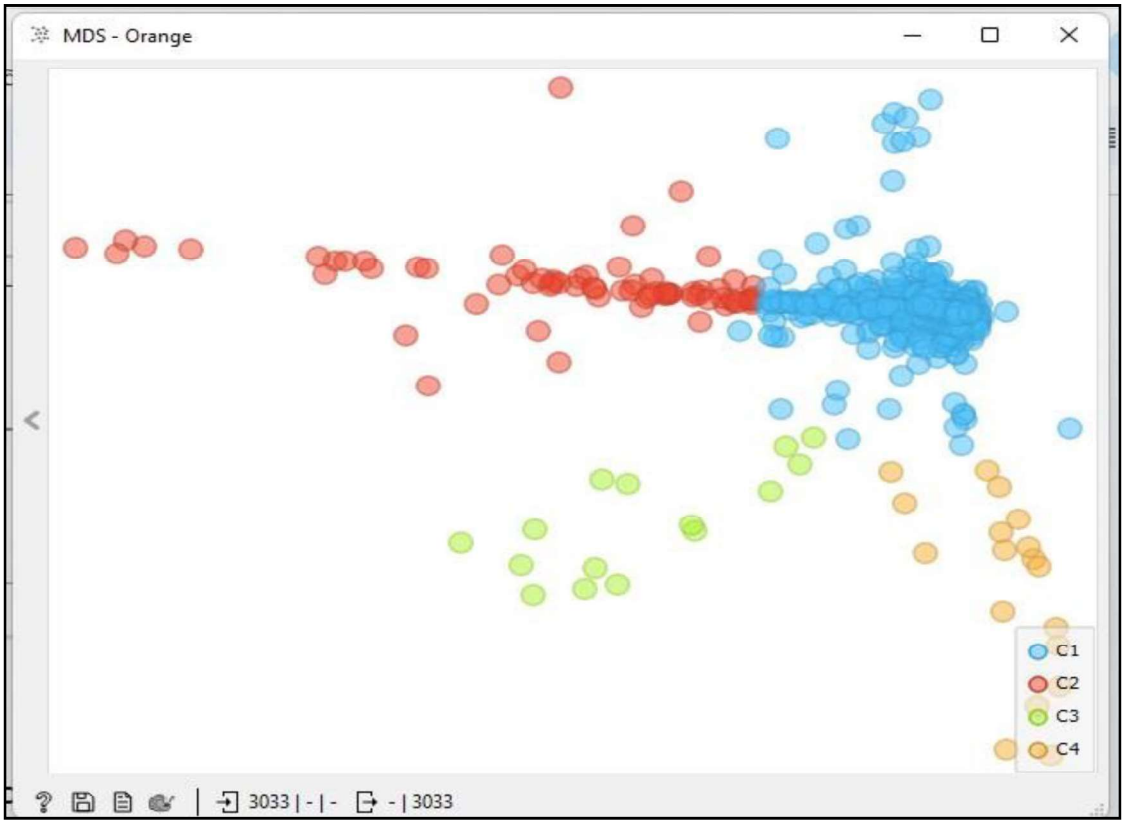
Fonte: Autora (2023).

A técnica de agrupamento hierárquico com a formação do dendograma não foi considerada satisfatória em função da quantidade de variáveis. Quando a quantidade de variáveis é grande o algoritmo *K-means* funciona de maneira melhor (SETIAWAN; KURNIAWAN; CHOWANDA; SUHARTONO, 2023).

Os vídeos tutoriais oficiais do Orange Data Mining ensinam algumas formas de se analisar e taxar os agrupamentos formados. Uma delas se resume em analisar os valores de “Student” por meio do gráfico *Box Plot*. A análise ocorre entre os valores do agrupamento referenciado em relação aos outros agrupamentos e suas variáveis. Em função da elevada quantidade de variáveis e a mensagem apresentada em relação a isso e os resultados do dendrograma, optou-se por extrair a informação das principais variáveis atribuídas ao agrupamento. As análises das variáveis, que permitiram caracterizar os grupos de códigos, foram realizadas pela informação emanado pelos gráficos de distribuição.

A figura 21 permite que se possa visualizar os agrupamentos em escala multidimensional. Esse resultado foi encontrado utilizando a função de escala multidimensional do programa e seria, apenas, mais uma forma de se representar os agrupamentos formados.

Figura 21 - Agrupamentos em escala multidimensional

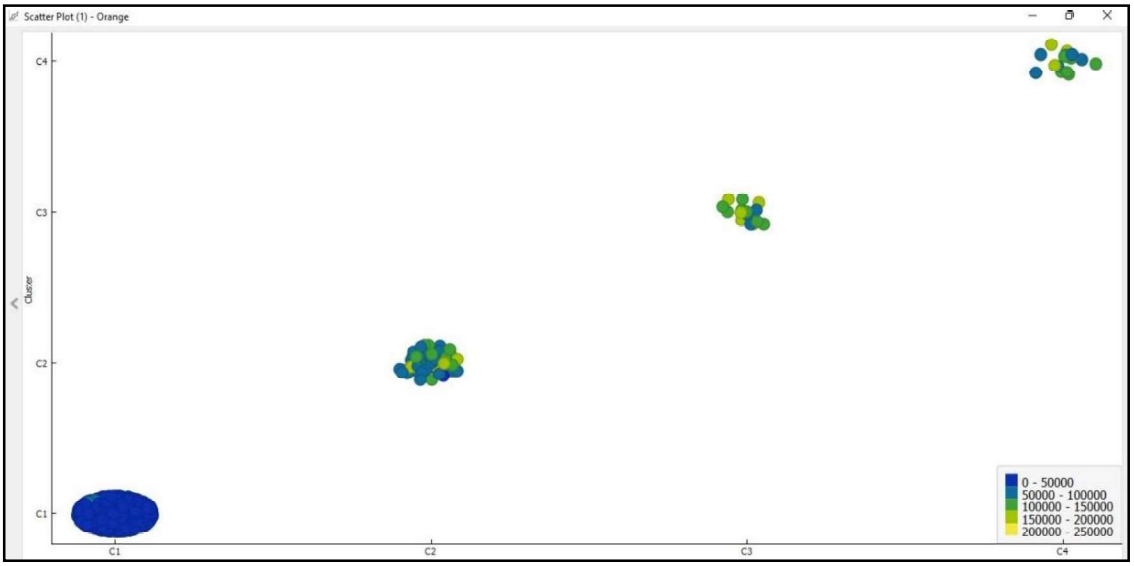


Fonte: Autora (2023).

As figuras 22, 23 e 24 caracterizam as variáveis custo, idade e sexo de acordo com os agrupamentos. Eles apresentam essas variáveis de forma bem segregada de maneira que não se pode inferir que os códigos foram agrupados de acordo com essas características.

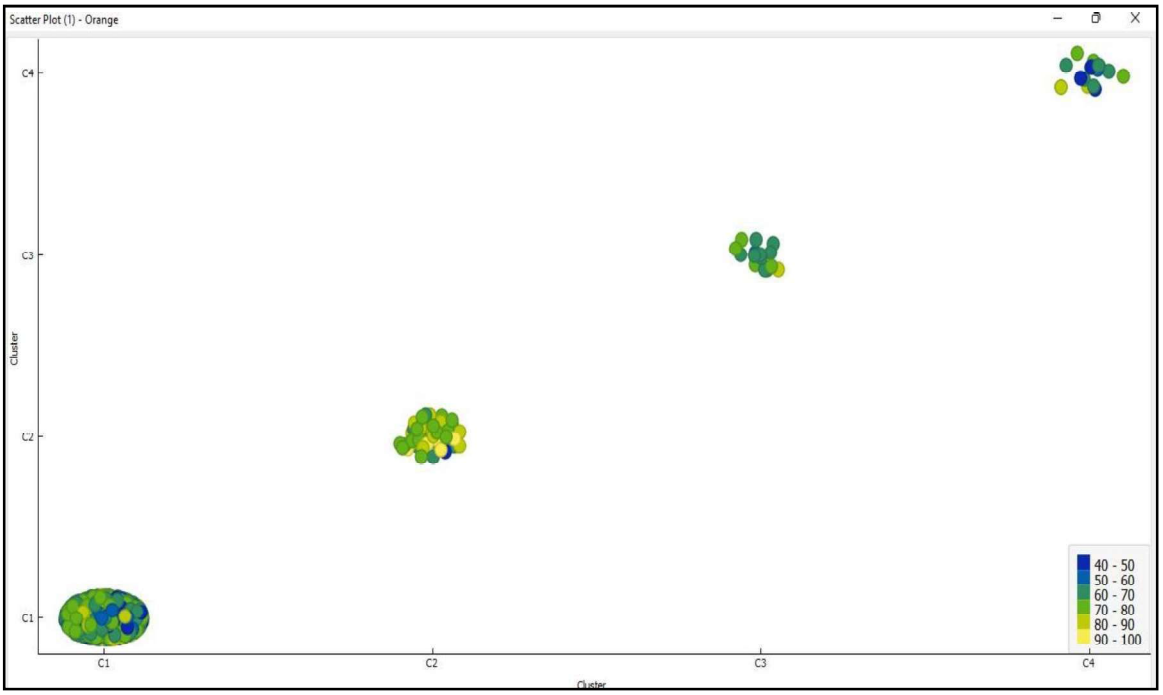
Observa-se, apenas, que o agrupamento 1 apresenta, em relação ao custo, a cor azul predominante de valores abaixo de R$5.000,00 e de códigos do sexo feminino.

Figura 22 - Representação dos agrupamentos pelo custo de utilização



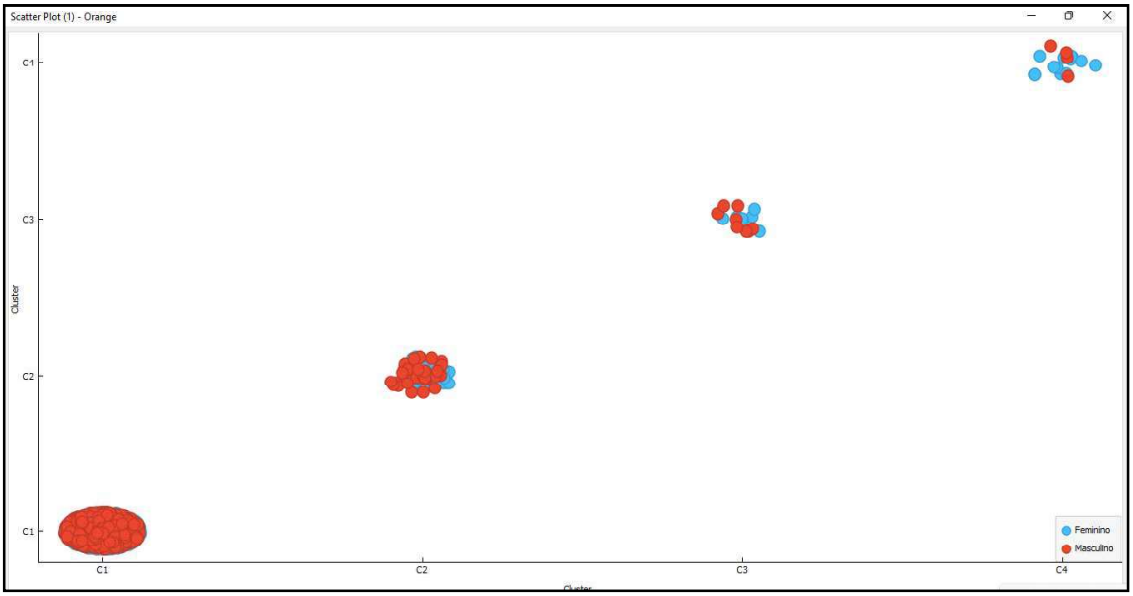
Fonte: Autora (2023).

Figura 23 - Representação dos agrupamentos pela idade dos usuários



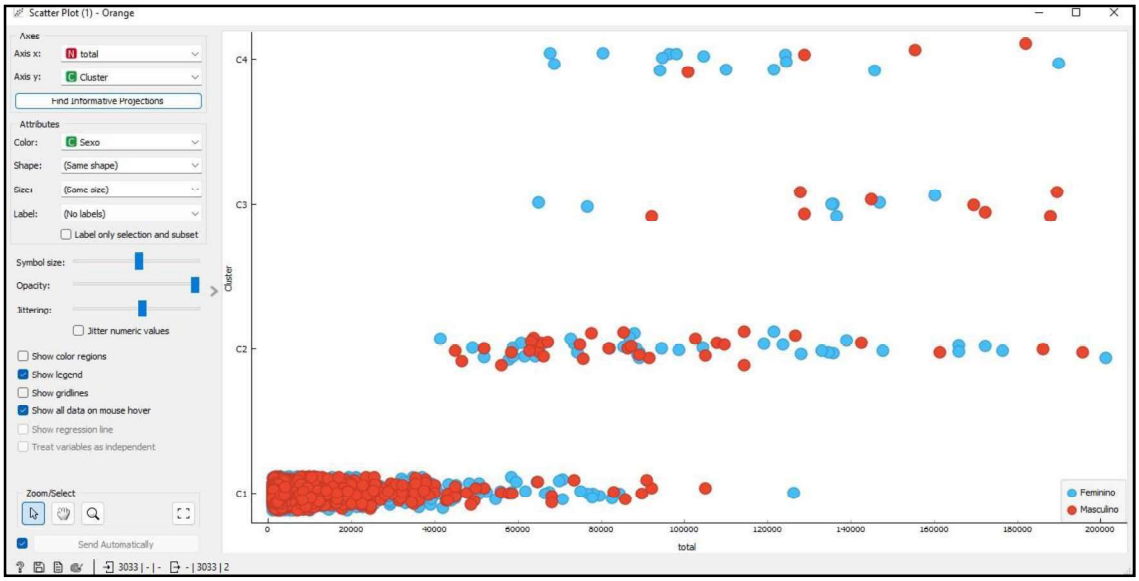
Fonte: Autora (2023).

Figura 24 - Representação dos agrupamentos pelo sexo dos usuários



Fonte: Autora (2023).

Figura 25 - Variação dos custos por agrupamento por idade

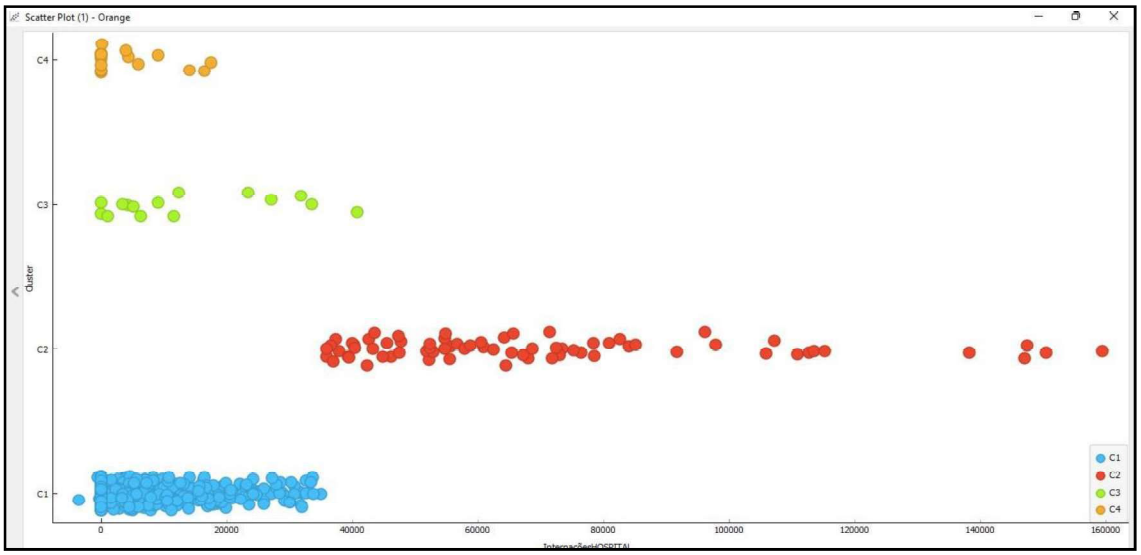


Fonte: Autora (2023).

Na figura 25 visualiza-se a diferença entre os custos dos usuários agrupados nos agrupamentos. As 4 segregações comportam usuários que gastaram entre R$60.000,00 e R$100.000,00, mas somente o agrupamento 1 agrupa usuários com gastos inferiores a R$4.000,00. Além disso, ele não possui nenhum código com valor maior de cerca de R$130.000,00.

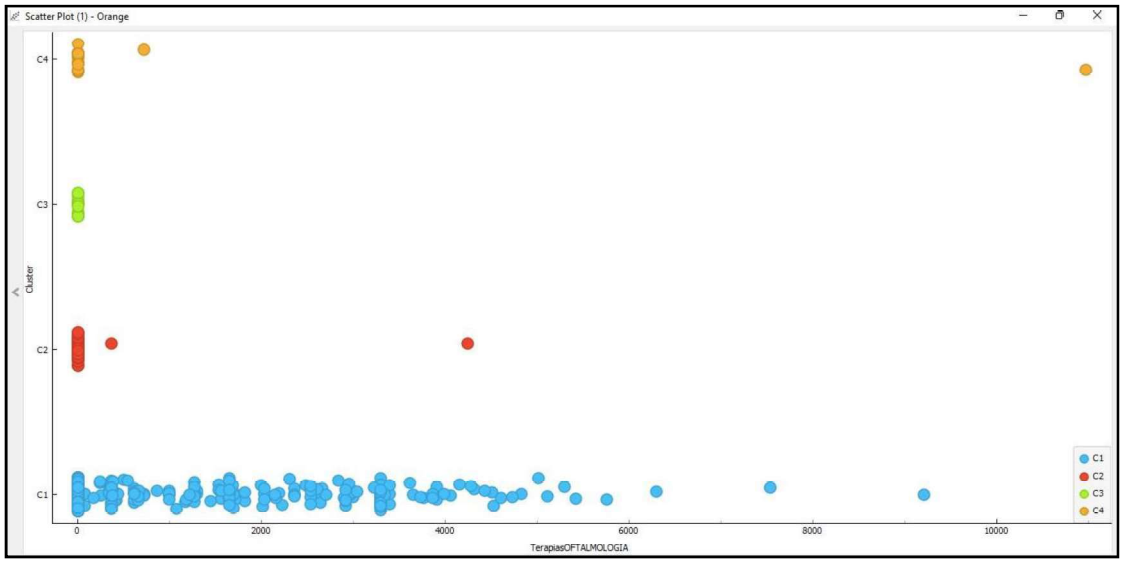
Na análise das características de cada segmentação podemos perceber, figura 26, que os maiores custos de internação hospitalar pertencem aos códigos juntados no agrupamento 2. De maneira igualmente bem distinta, podemos verificar que os códigos com custos de procedimentos da especialidade de oftalmologia estão unidos no agrupamento 1, figura 27.

Figura 26 - Internação Hospitalar por custo separada por agrupamento



Fonte: Autora (2023).

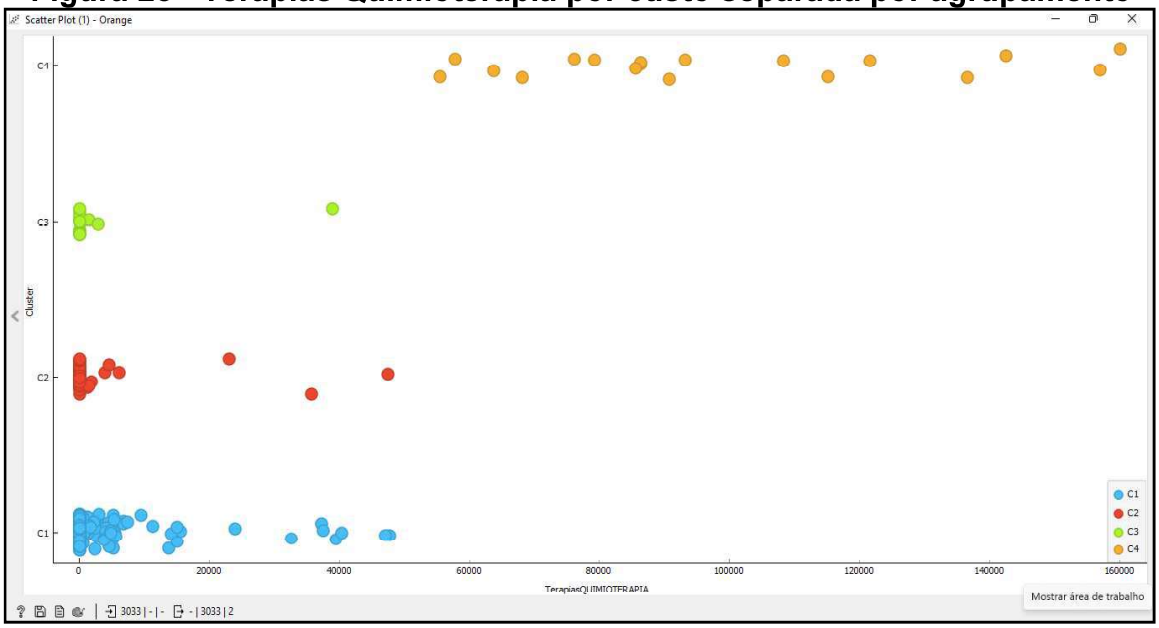
Figura 27 - Terapias Oftalmologia por custo separada por agrupamento



Fonte: Autora (2023).

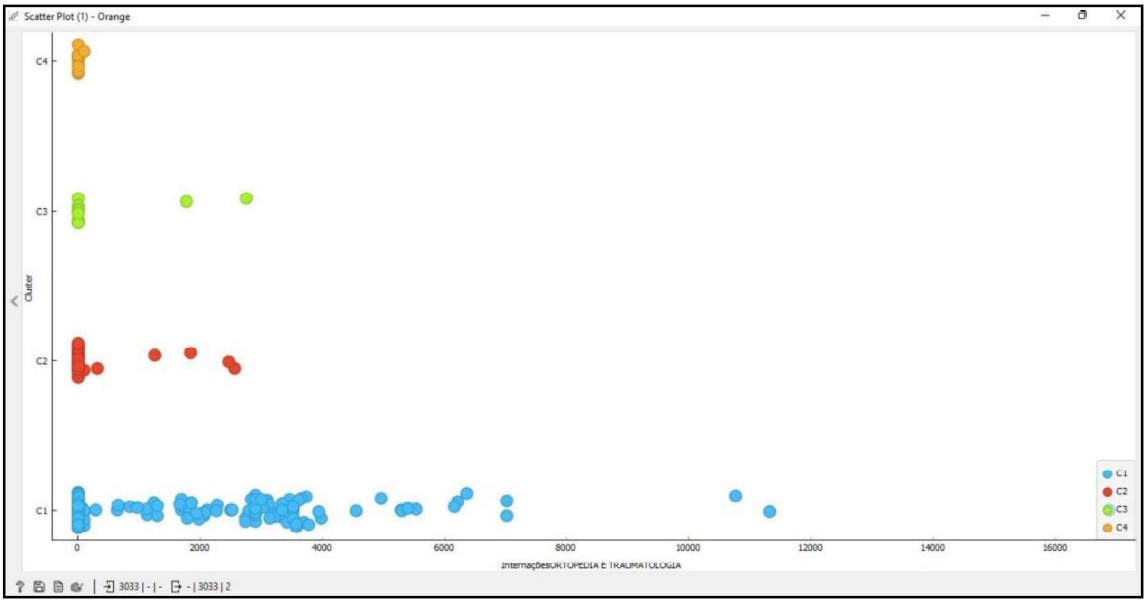
Da figura 28 se retira a informação de que no agrupamento 4 estão agrupados os usuários que fazem uso de medicamentos quimioterápicos de alto custo e que a maioria das pessoas que realizou algum procedimento relacionado à internação relacionados à traumas e à ortopedia fazem parte do agrupamento 1 (figura 29 e 30).

Figura 28 - Terapias Quimioterapia por custo separada por agrupamento



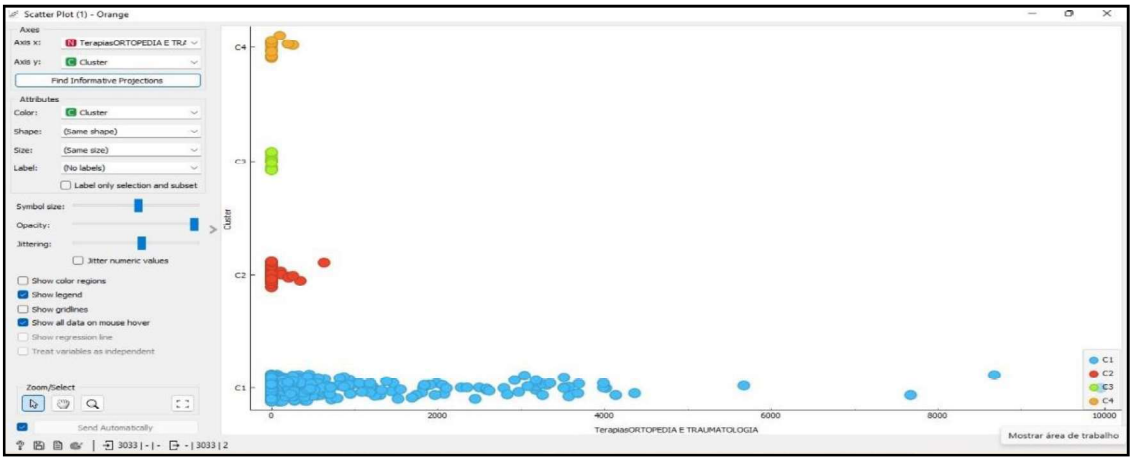
Fonte: Autora (2023).

Figura 29 - Internações Ortopedia e traumatologia por custo separada por agrupamento



Fonte: Autora (2023).

Figura 30 - Terapia ortopedia e traumatologia separada por agrupamento e valores



Fonte: Autora (2023).

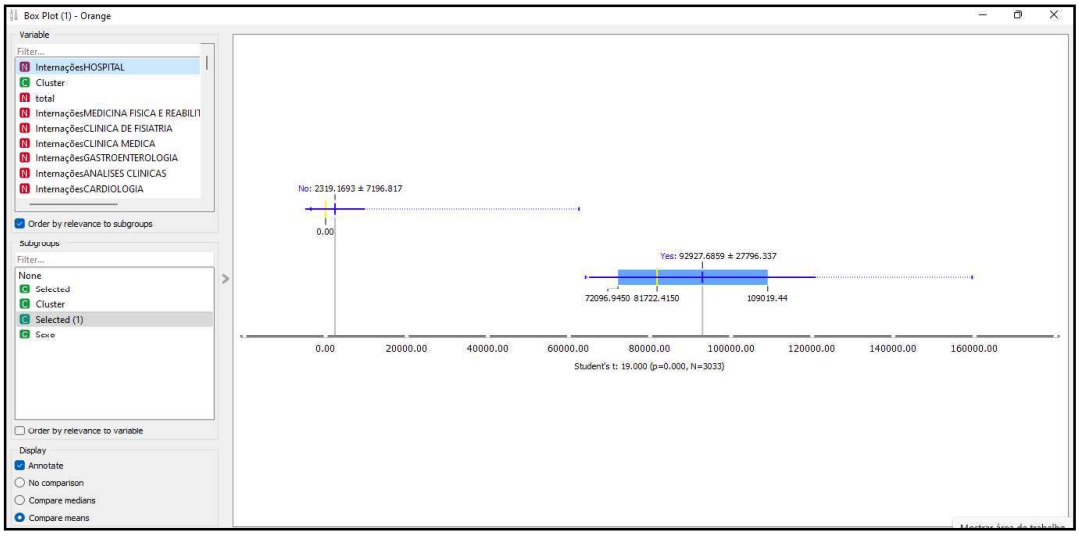
Nos próximos parágrafos serão analisadas as características singulares de cada agrupamento.

6.2 ANÁLISE DOS AGRUPAMENTOS

6.2.1 Agrupamento 1 – Pacientes com condições simples de saúde

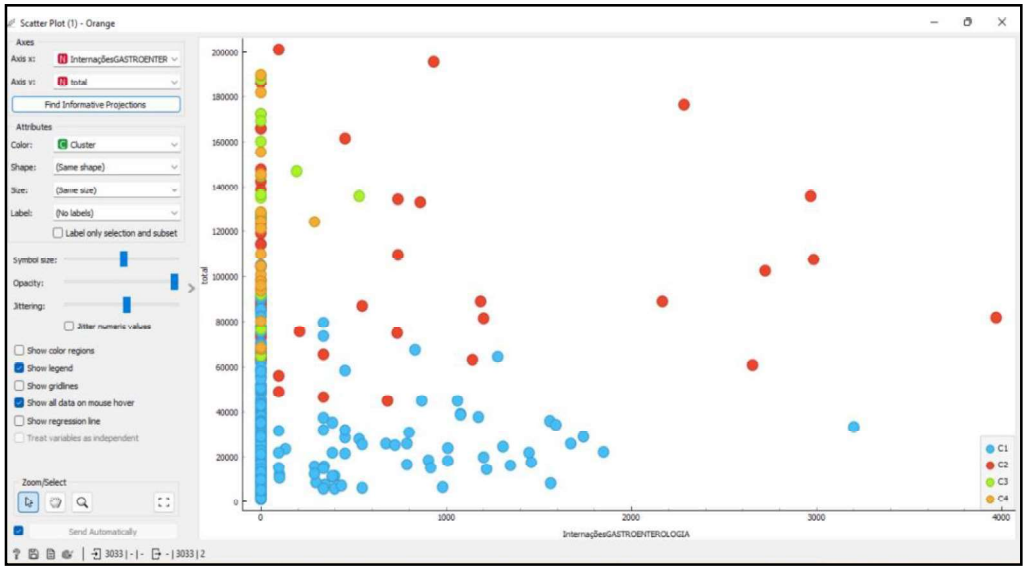
O agrupamento 1 é composto de 2929 usuários. As principais variáveis atribuídas a ele constam na figura 31.

Figura 31 - Principais variáveis atribuídas ao agrupamento 1



Fonte: Autora (2023).

Figura 32 - Internações Gastroenterologia por custo e agrupamento



Fonte: Autora (2023).

Podemos evidenciar que o agrupamento 1 possui mais usuários que fizeram uso de procedimentos relacionados a internações referentes à gastroenterologia, figura 32. Essa classificação se deve mais à quantidade de usuários que ao valor despendido.

Todos os pacientes agrupados nesta classificação tiveram custo com internação hospitalar menor de R$40.000,00 (quarenta mil reais) – figura 26.

Depreende-se da análise de todas as variáveis que neste grupo foram incluídos usuários que utilizaram o plano de maneira menos onerosa. A coluna de frequência de utilização/custo das variáveis apresenta comportamento parecido no sentido de a frequência ser alta, maior de 2800, e o custo com determinado grupo de procedimento ser menor de R$200,00.

Podemos demonstrar isso, não só pela figura 34, mas também pela figura 33 - Consultas médicas endocrinologia. Dos 2929 usuários, apenas 59 tiveram custos maiores de R$50,00 e menores de R$450,00 com consultas ao especialista de endocrinologia. 2866 usuários tiveram gastos de R$0,00 a R$50,00 nesta especialidade.

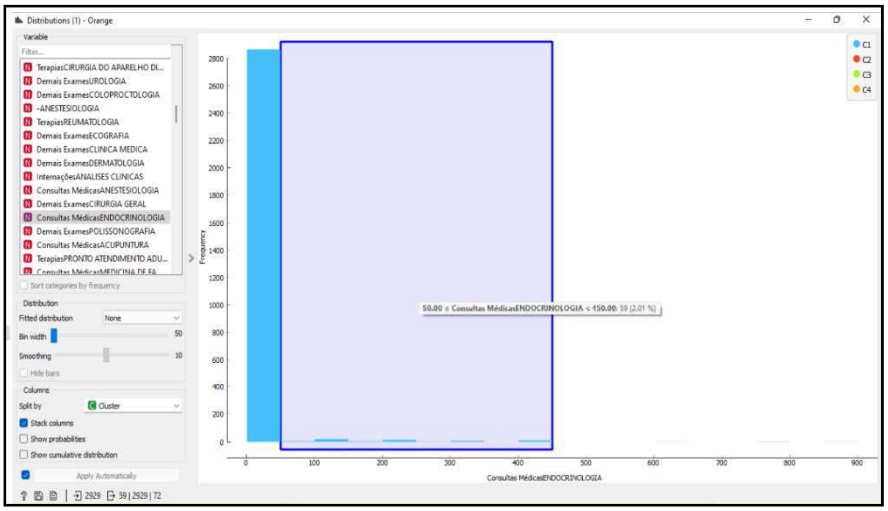
Pode-se observar um comportamento um pouco diferente no atributo da figura 35 - Consultas médicas ortopedia e traumatologia, mesmo assim não se trata de nenhum gasto exagerado se considerarmos que a carteira de usuários de uma maneira geral é idosa e o período de análise compreende 12 (doze) meses.

Analisando as variáveis relacionadas à tratamentos e consultas oftalmológicos

entende-se que poderiam representar um grupo de usuários que utilizou o plano de maneira diferenciada e que estão agrupados neste agrupamento – figura 37. 1480, ou seja, 50,53% dos usuários classificados neste agrupamento tiveram custo com essa especialidade de R$0,00 a R$50,00 e 49,47% dos usuários gastaram entre R$50,00 e R$1.600,00.

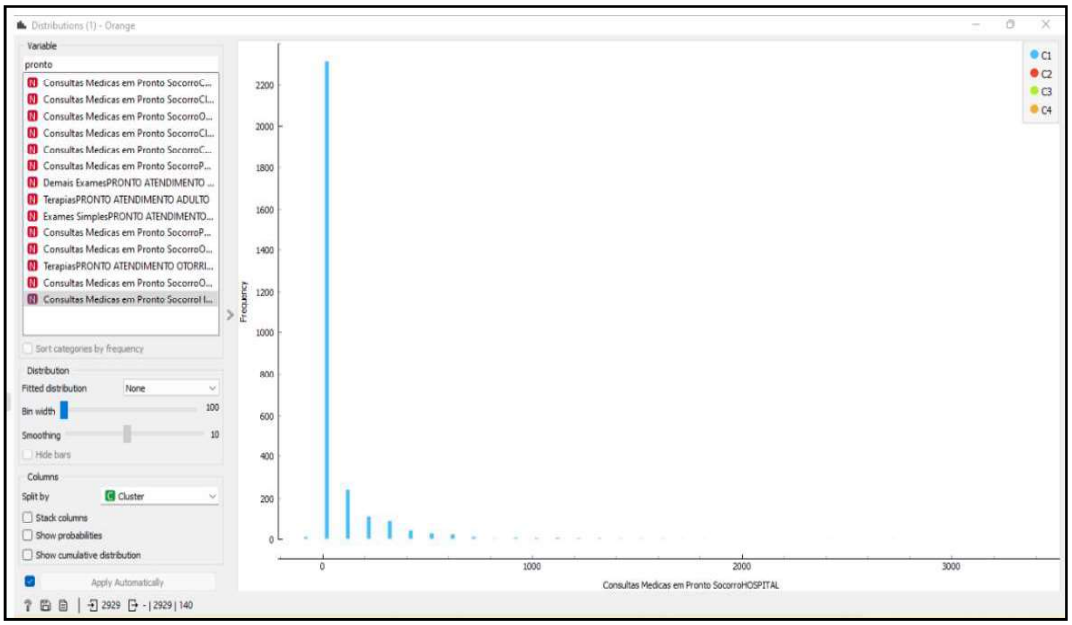
Pode-se notar algum comportamento que sinaliza o uso um pouco mais acentuado nas especialidades de ginecologia e obstetrícia e nas consultas médicas em pronto socorro – figuras 34 e 36.

Figura 33 - Consultas médicas endocrinologia



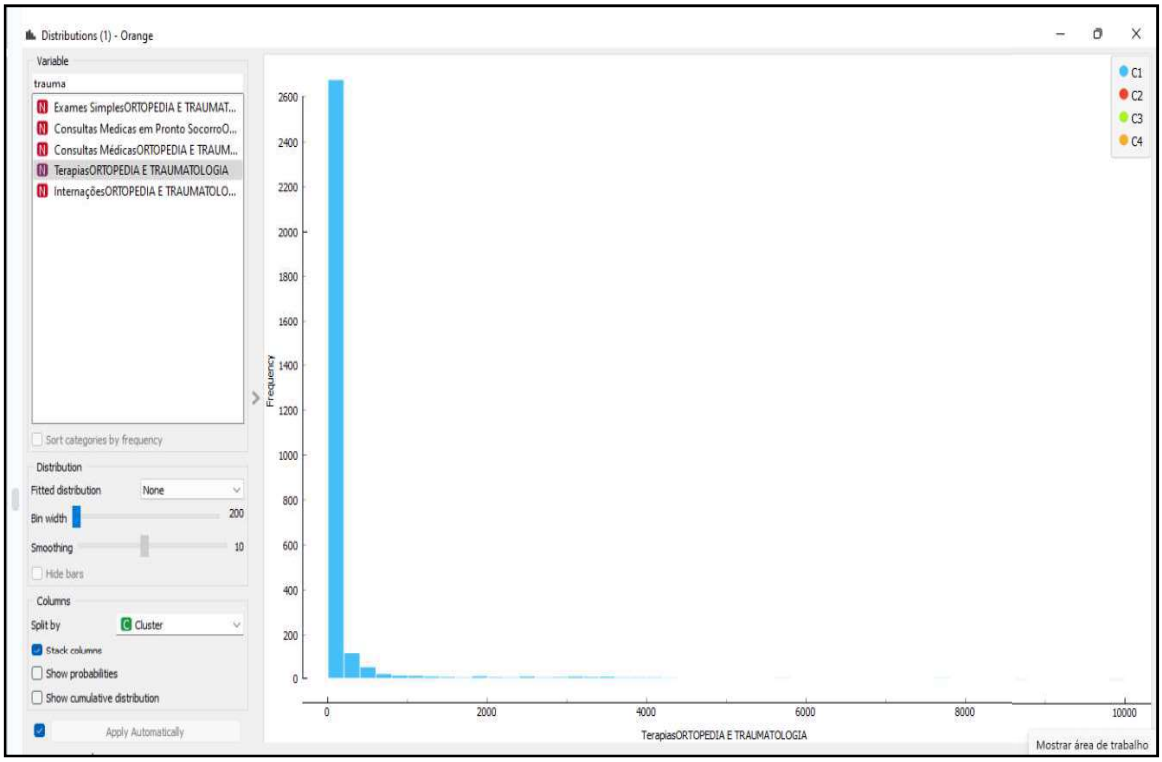
Fonte: Autora (2023).

Figura 34 - Consultas médicas em Pronto Socorro



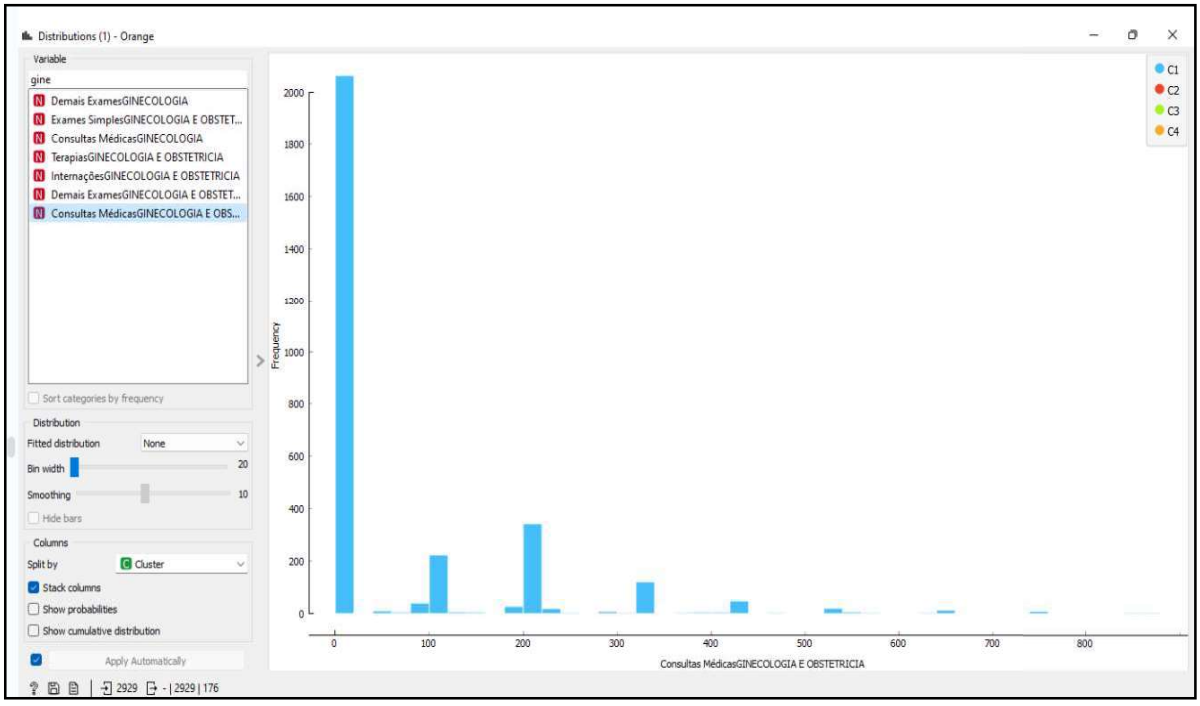
Fonte: Autora (2023).

Figura 35 - Terapias Ortopedia e Traumatologia



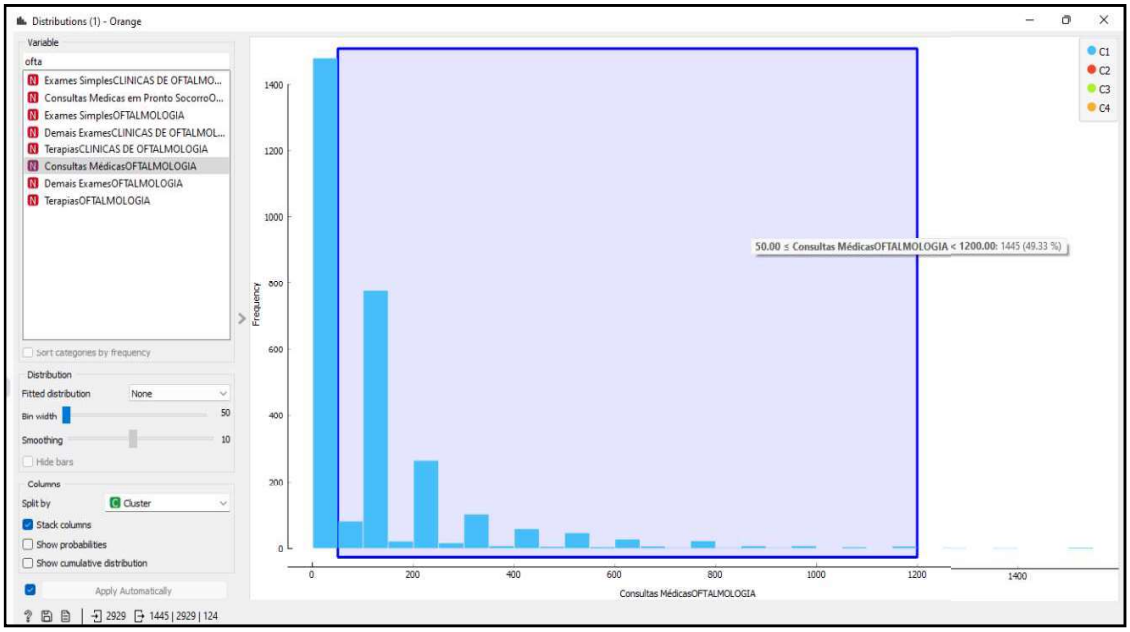
Fonte: Autora (2023).

Figura 36 - Consultas médicas Ginecologia e Obstetrícia



Fonte: Autora (2023).

Figura 37 - Consultas médicas Oftalmologia



Fonte: Autora (2023).

Esse agrupamento é o mais numeroso dos 4. Dos 2929 integrantes desse grupo, 1714, ou seja, 58,52%, tiveram gastos anuais com plano de saúde de até R$5.000,00. Destes, 1076 (72,75%) possuem idade entre 60 e 100 anos. Se considerarmos um ticket médio de R$1.000,00 mensais de gastos com plano de saúde, totalizaria R$12.000,00 anuais. A relação receita/despesa, ou seja, a sinistralidade desse grupo ficaria em 41,67%. Esse número é considerado baixo se formos analisar a sinistralidade média das operadoras de saúde em relação a essa faixa etária. Entende-se assim, que esses integrantes, não seriam os primeiros beneficiários de ações para melhoria de saúde ou para a redução de gastos.

Uma ação que poderia ser tomada e que beneficiaria todos os usuários desse agrupamento seria analisar a quantidade de consultas utilizadas por especialidade médica com o intuito de se encontrar usuários hiper utilizadores de consultas e que consultam, desnecessariamente, com especialistas. Isso pode acontecer com pacientes que não conseguem resolver seu problema de saúde ou não sabem qual especialista procurar num primeiro momento e acabam consultando com vários. Ações relacionadas a entrar em contato com esse paciente e divulgar opções relacionadas à atenção primária como apresentar um médico de família a ele. Esse profissional poderá orientar o paciente sobre as questões de saúde e a necessidade de se procurar um especialista. Essas ações organizam os usuários dentro do sistema

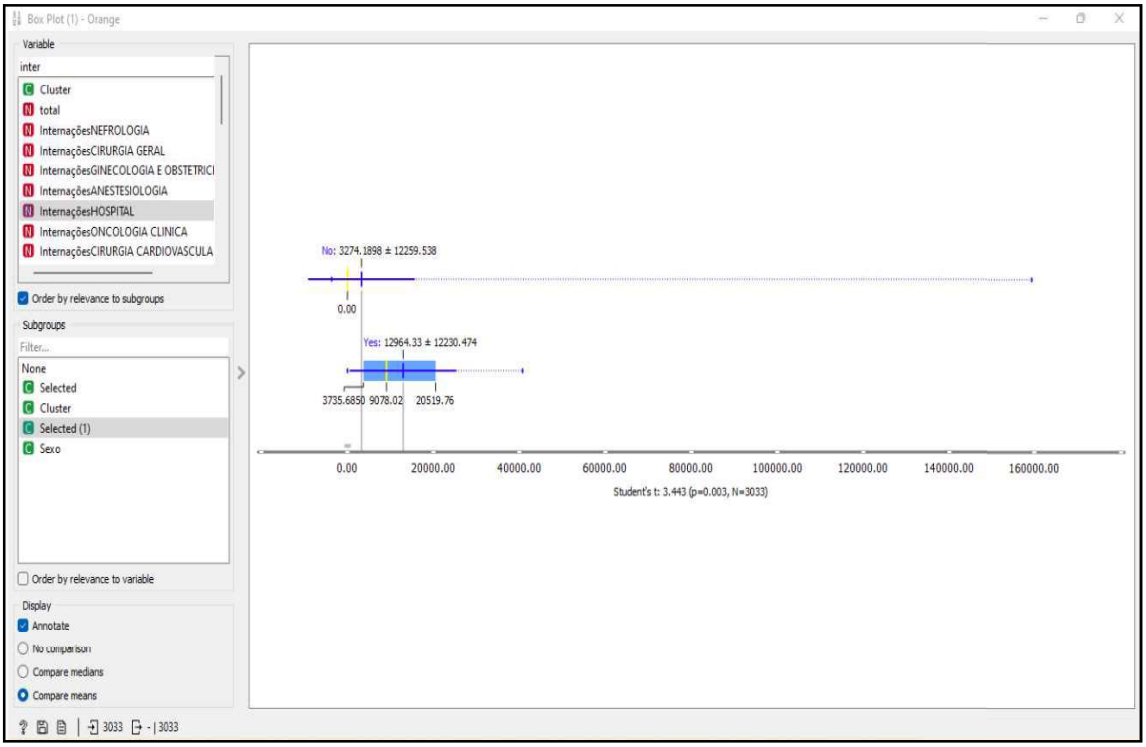
de saúde para que usem seu serviço de maneira adequada e assertiva (PIRES et. al, 2019).

Há a possibilidade, ainda, de se selecionar alguns pacientes para acompanhamento a longo prazo e analisar, temporalmente, sua evolução de gastos. Isso seria importante para mapear gastos presentes e futuros e selecionar procedimentos que poderiam ser variáveis num modelo preditivo de aprendizado de máquina supervisionado.

6.2.2 Agrupamento 2 – Pacientes com altos custos hospitalares

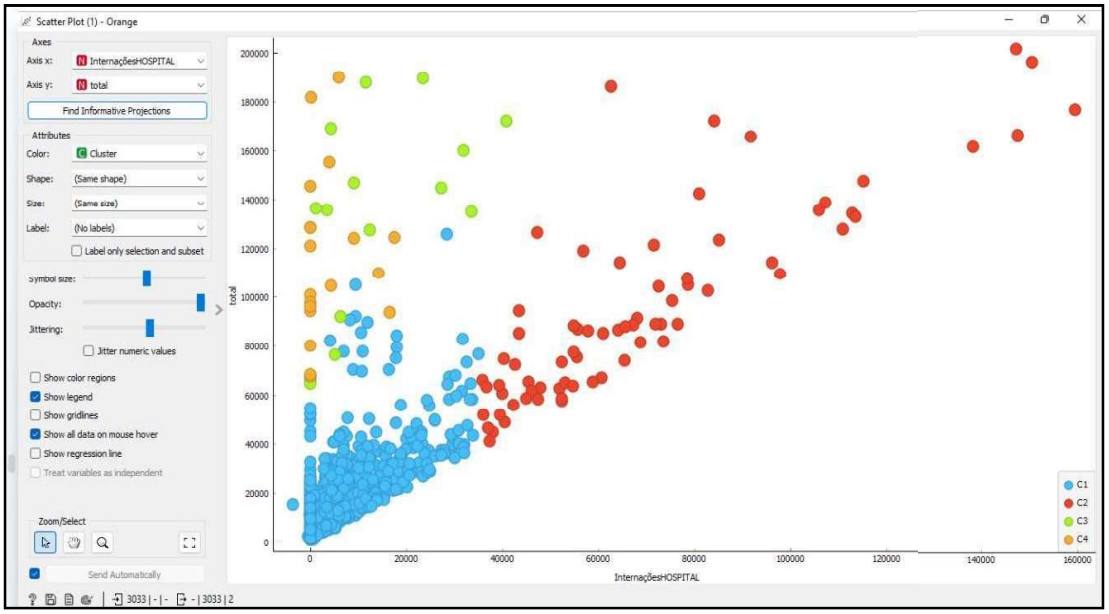
O agrupamento 2 é composto de 71 usuários e suas principais variáveis constam na figura 38.

Figura 38 - Principais variáveis atribuídas ao agrupamento 2



Fonte: Autora (2023).

Figura 39 - Internações Hospital por custo e agrupamento

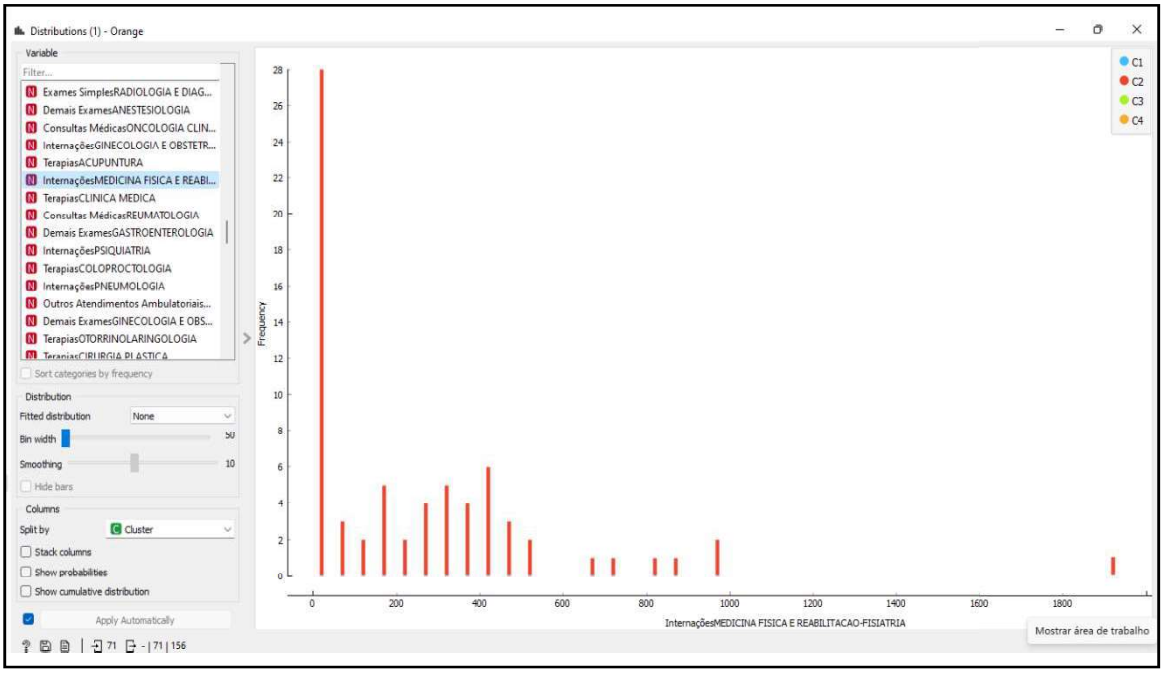


Fonte: Autora (2023).

A figura 39 demonstra com nitidez a importância do atributo “Internações Hospital” na caracterização do agrupamento 2. Nele constam os pacientes com gastos de internações hospitalares maiores de R$40.000,00 (quarenta mil reais).

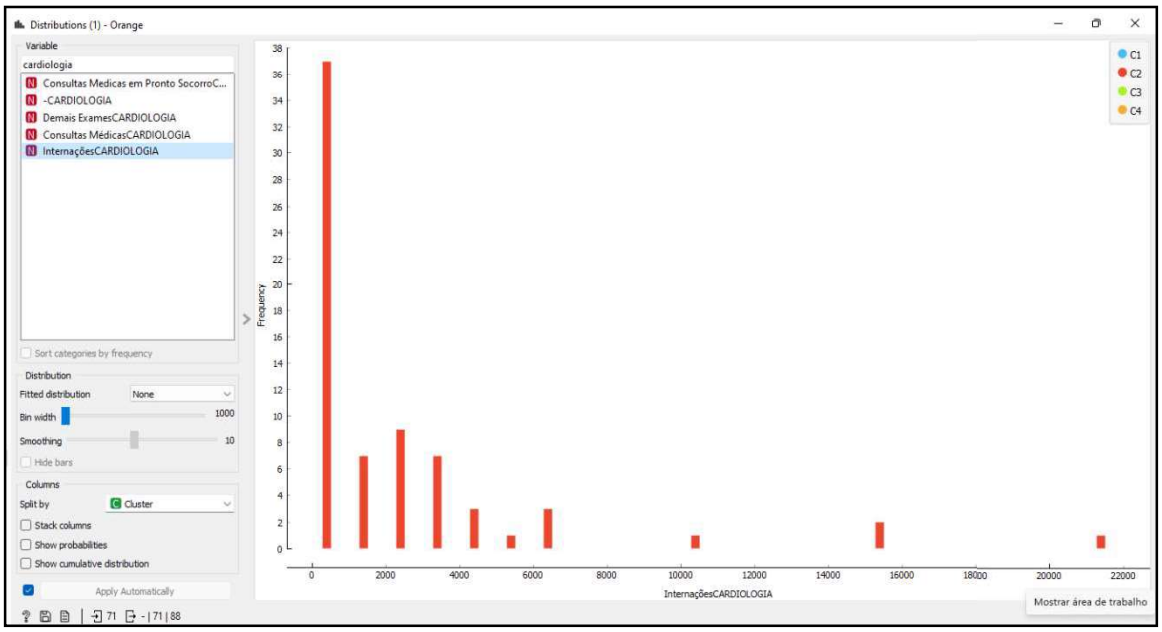
Podemos relacionar a esse agrupamento maiores custos com medicina de reabilitação – figura 40, e gastroenterologia (figura 38) além de maiores valores relacionadas às internações neurologia e neurocirurgia, internações cirurgias cardiovasculares (figura 41) e exames cardiológicos, consultas médicas em pronto socorro (figura 43) e terapias cirurgia geral e clínica médica (figura 42), internações urologia e internações nefrologia. Também foi selecionado o gráfico que explicita o resultado do atributo sobre serviços de imagem (figura 44) e de internações em anestesiologia (figura 45).

Figura 40 - Internações Medicina Física e reabilitação



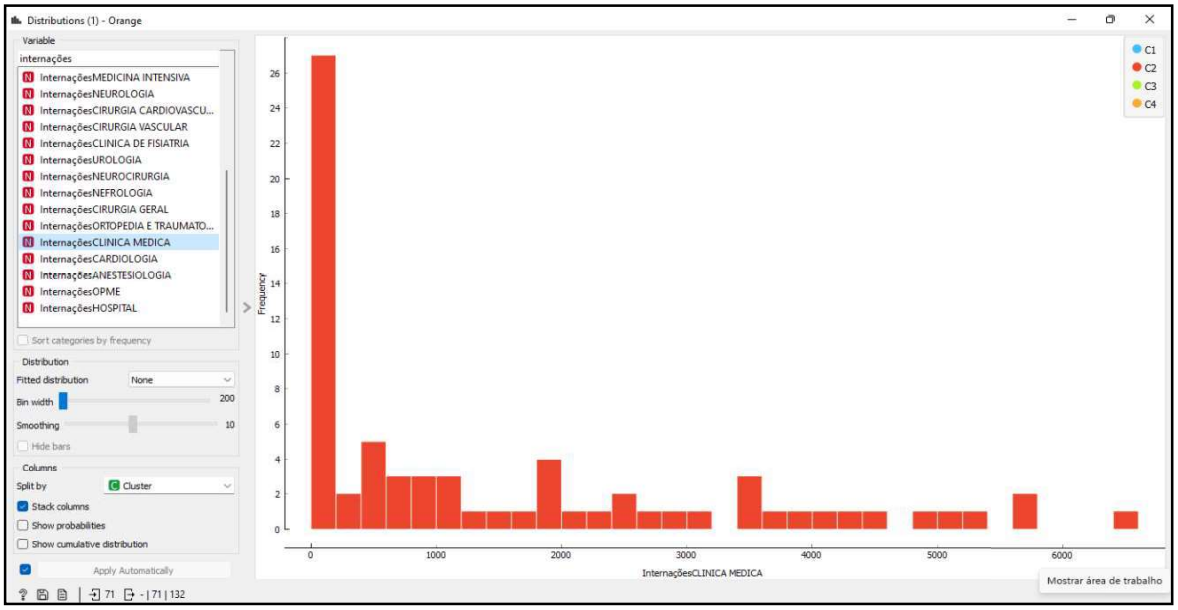
Fonte: Autora (2023).

Figura 41 Internações cardiologia



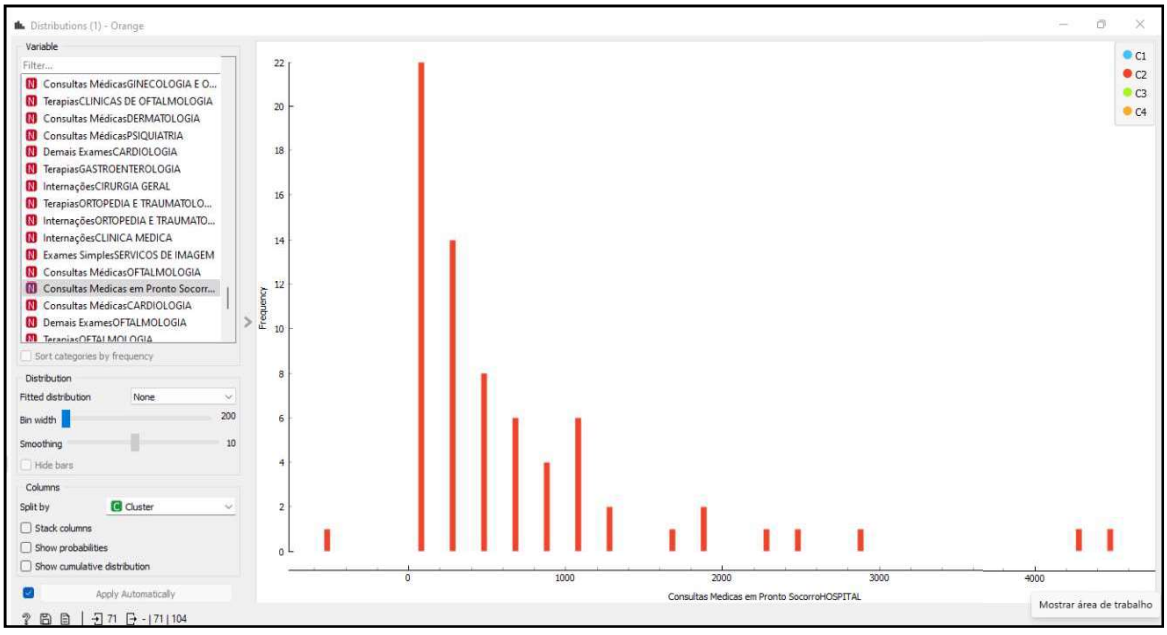
Fonte: Autora (2023).

Figura 42 - Internações clínica médica



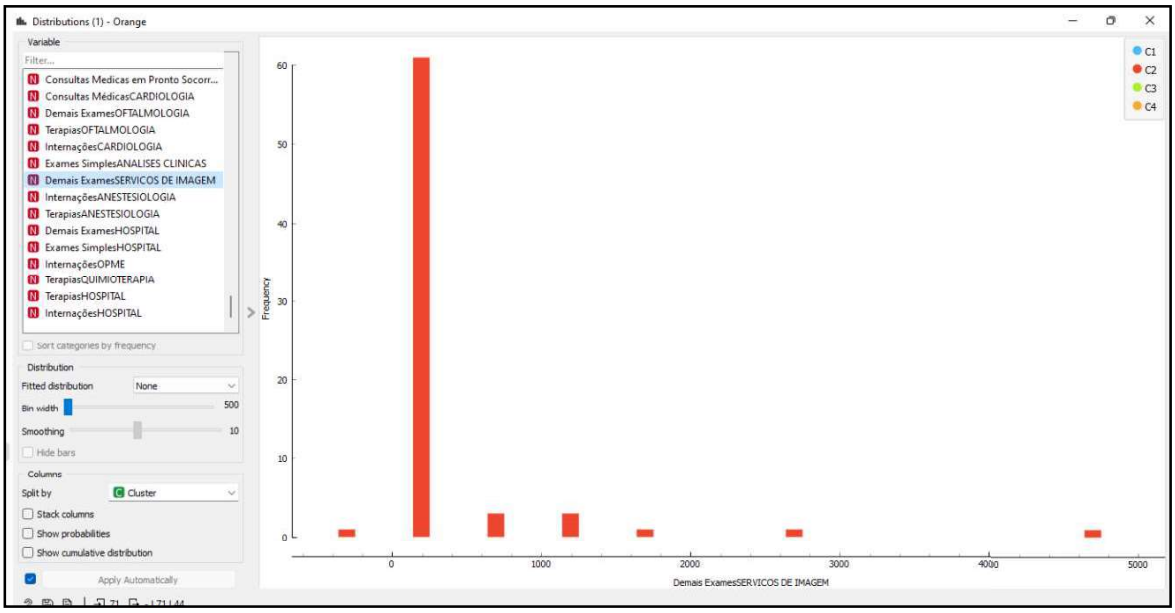
Fonte: Autora (2023).

Figura 43 - Consultas médicas em pronto socorro



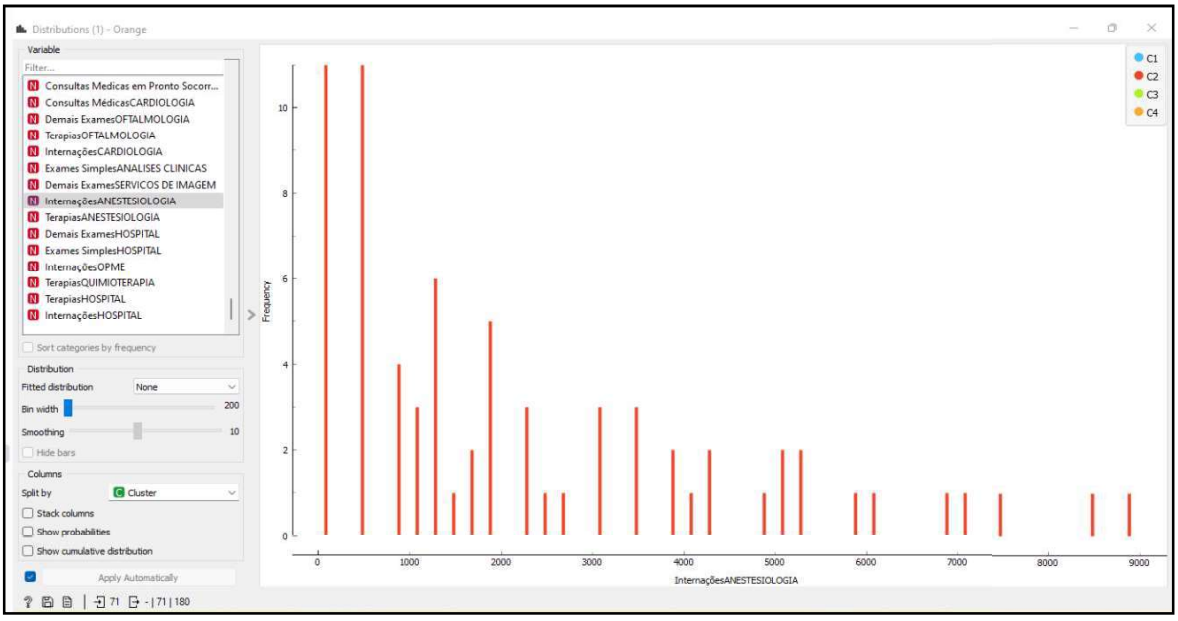
Fonte: Autora (2023).

Figura 44 - Exames serviços de imagem



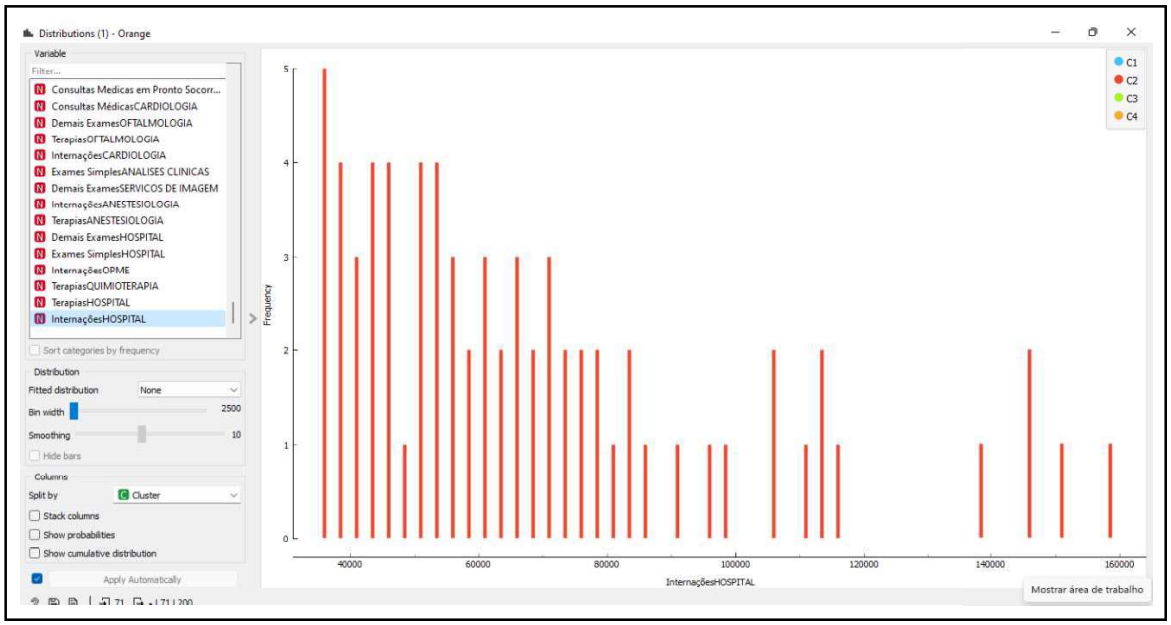
Fonte: Autora (2023).

Figura 45 - Internações anestesiologia



Fonte: Autora (2023).

Figura 46 - Internações hospital



Fonte: Autora (2023).

Este agrupamento apresenta, em sua maioria, os pacientes que tiveram altos custos com internações.

Como ações em saúde que beneficiariam esse grupo citam-se o contato com esses pacientes para entender o motivo das cirurgias e oferecer ajuda em cuidados em alta hospitalar. O contato pós internação é importante para acompanhar a saúde do usuário de forma ampla e tomar conhecimento sobre reinternações.

Acompanhar pacientes que tiveram vários períodos de internação em um curto espaço de tempo se torna valioso à medida em que se pode, por meio dos relatórios de custo, descobrir conhecimento a respeito dos procedimentos recorrentes em pacientes que tiveram essa condição. Assim, poderia ser aplicado algum modelo de aprendizado de máquina supervisionado a eles a fim de predizer pacientes sujeitos a esses eventos e, por fim, tentar evitar ou minimizar a ocorrência e o desgaste para o paciente e sua família (YUILL; KUNZ, 2022)

Um aspecto importante em relação à saúde do paciente, ao atendimento hospitalar e aos custos relacionados às internações é o acompanhamento desse paciente em relação às infecções pós-operatórias e o desfecho causado por esse acontecimento. Infecções hospitalares aumentam o período de estadia no hospital e, consequentemente, o desconforto para o paciente e seus familiares. Acompanhar esses casos é importante para prestar apoio ao paciente, para assimilar casos recorrentes e sinalizar à auditoria do hospital e para entender que tipo de população

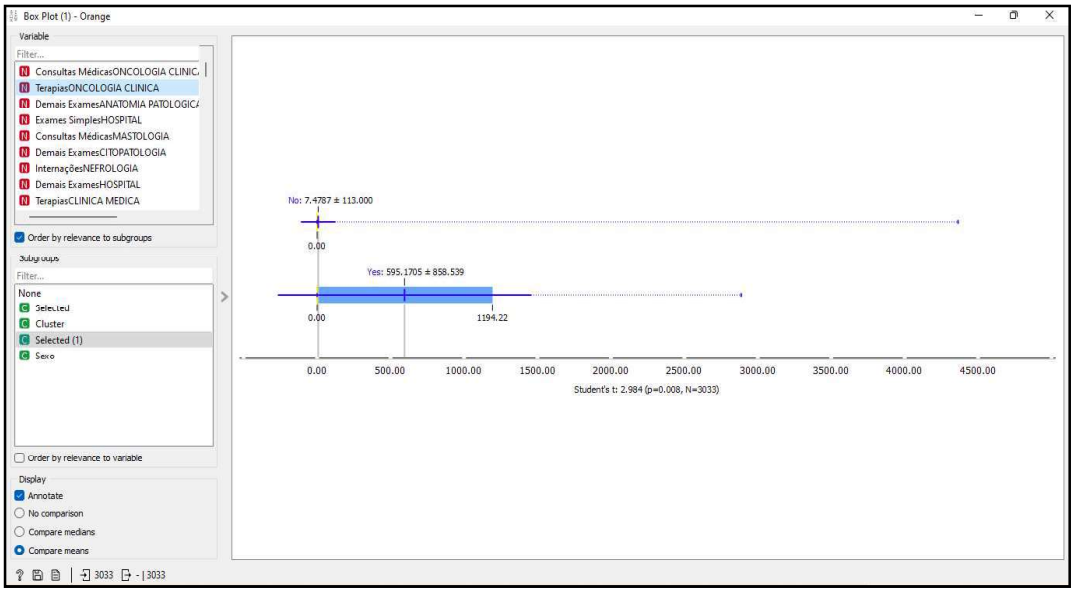
está sendo atingida e em qual local isso tem acontecido.

Entende-se que os associados com maiores gastos nas especialidades de cardiologia, de endocrinologia e de nefrologia estão agrupados nos agrupamentos 1 e 2. A busca ativa por usuários que utilizam procedimentos de forma acentuada nessas especialidades se faz necessária para encontrar pacientes de doenças crônicas como a hipertensão, a diabetes e a insuficiência cardíaca. O contato com esses pacientes no sentido de monitorar sua condição de saúde é uma boa estratégia para aumentar o bem-estar dessas pessoas e para ajudá-los no controle da doença. Um paciente diabético precisa estar atento aos seus níveis de glicose sanguínea assim como um hipertenso em relação à sua pressão arterial. Essas ações são válidas no sentido de evitar que esses pacientes descompensem e passem a usar o sistema de saúde de forma onerosa (FONSECA; OGATA, 2021).

6.2.3 Agrupamento 3 – Pacientes oncológicos com internações

O agrupamento 3 é composto por 15 usuários e suas variáveis principais constam na figura 47.

Figura 47 - Principais variáveis atribuídas ao agrupamento 3



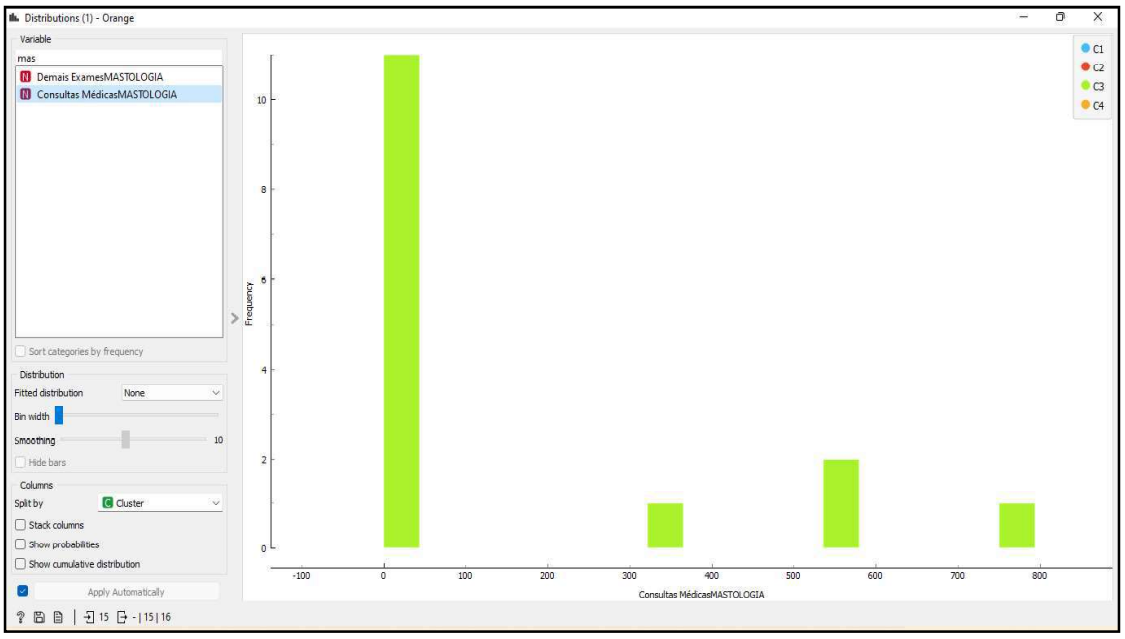
Fonte: Autora (2023).

Depreende-se da análise das variáveis com melhores resultados, neste agrupamento, que nele estão agrupados pacientes submetidos a tratamentos

oncológicos que tiveram internações relacionadas ao tratamento e à doença (figura 47 e 50).

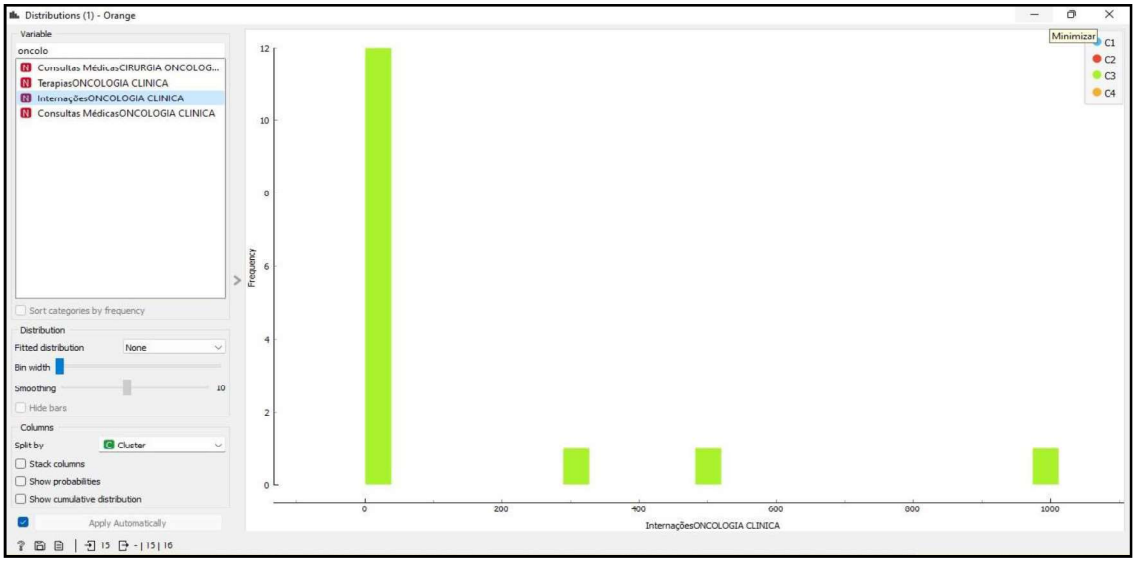
Encontraram-se maiores valores associados a especialidade de mastologia (figura 48) assim como frequência elevada de procedimentos nas variáveis relacionadas a consultas médicas em pronto socorro, o que ficaria ao encontro da condição de pacientes oncológicos que permaneceram algum tempo internados. Além disso, foram analisadas frequências acentuadas (comparadas à quantidade de usuários) das variáveis que retratam procedimentos associados a consultas médicas em radioterapia, exames de anatomia patológica, consultas médicas cirurgia oncológica e internação (figura 49), consultas médicas em medicina intensiva, exames de diagnóstico por imagem, consultas nefrologista, fisioterapia hospitalar, exames e consultas referentes à pneumologia e exames cardiológicos.

Figura 48 - Consultas médicas Mastologia do agrupamento 3



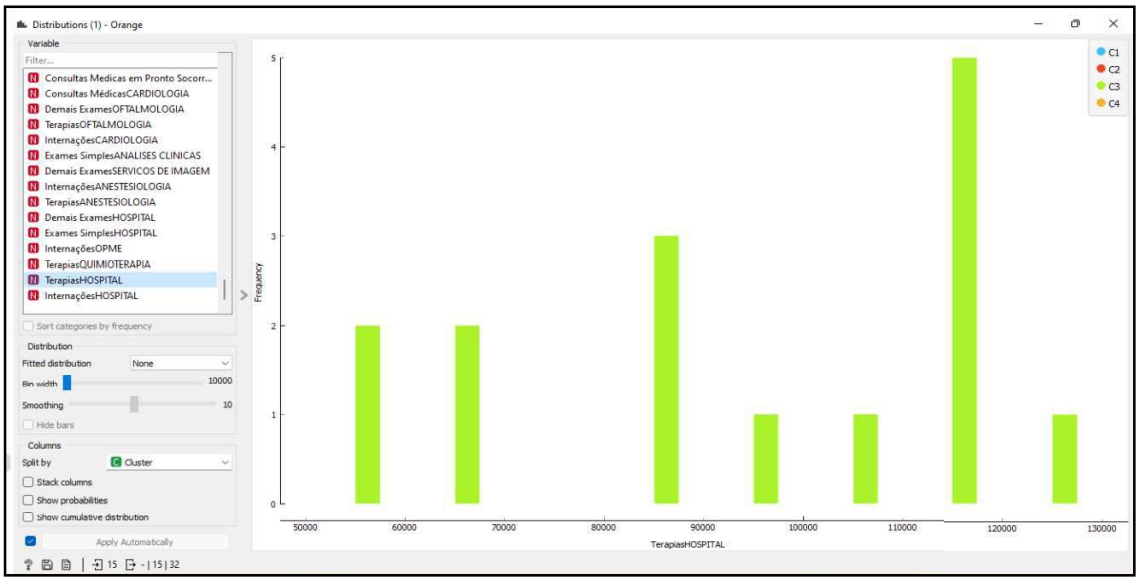
Fonte: Autora (2023).

Figura 49 - Internações Oncologia Clínica do agrupamento 3



Fonte: Autora (2023).

Figura 50 - Terapias hospital do agrupamento 3



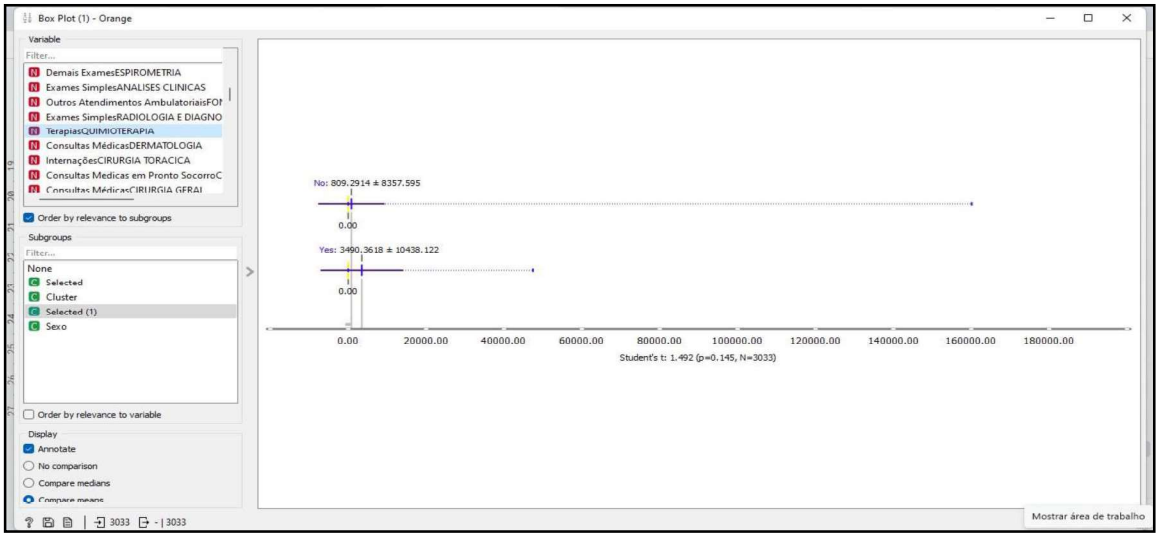
Fonte: Autora (2023).

Uma forma de se entender o custo desses pacientes e suas doenças a fim de se implantar algum tipo de ação que beneficia esse grupo e potenciais portadores de algum tipo de câncer seria acompanhar esses pacientes por mais de um ano. Assim, seria possível verificar se eles fizeram algum tratamento oncológico anteriormente à sua internação e se seguiram os protocolos de exames e de consultas a profissionais da saúde de acordo com as suas condições de saúde pregressas e as suas idades.

6.2.4 Agrupamento 4 – Pacientes oncológicos

O agrupamento 4 é composto de 18 usuários e suas variáveis principais constam na figura 51.

Figura 51 - Principais variáveis atribuídas ao agrupamento 4

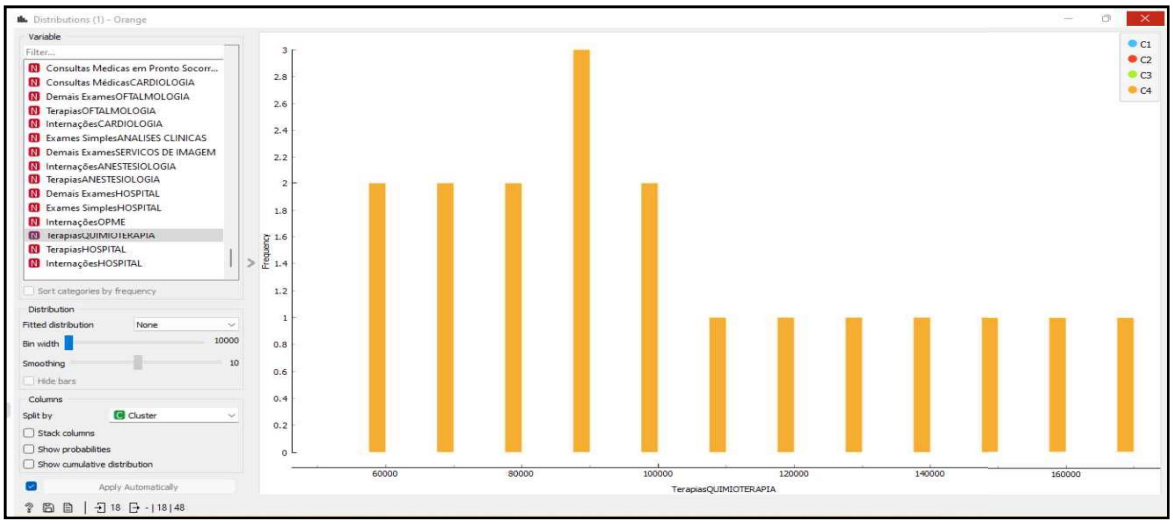


Fonte: Autora (2023).

Neste grupo estão reunidos os usuários do plano de saúde que tiveram altos custos (acima de 60 mil) com medicamentos quimioterápicos – figura 52 - e baixos custos de internações.

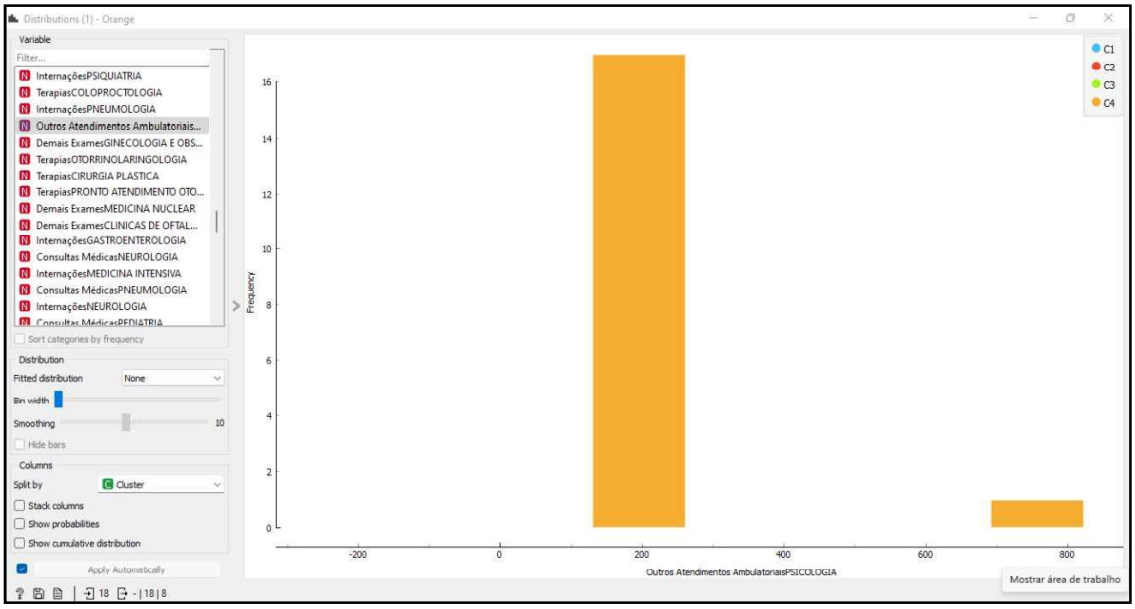
Observa-se uma frequência elevadas de consultas de psicologia (figura 53).

Figura 52 - Terapias quimioterapia



Fonte: Autora (2023).

Figura 53 - Consultas psicologia



Fonte: Autora (2023).

Este agrupamento se mostra relevante por selecionar os pacientes que fazem uso de medicamentos quimioterápicos. Estes medicamentos, em sua maioria, são caracterizados como de alto custo.

Como forma de gerir esse grupo e salientar a importância deste agrupamento podemos citar o acompanhamento desses pacientes para o aumento ou diminuição do tratamento e se os pacientes evoluirão para algum tratamento cirúrgico ou para a necessidade de internação hospitalar.

Interessante, também, seria selecionar quais tipos de câncer os usuários estão tratando e quais quimioterápicos estão sendo utilizados.

Os agrupamentos 3 e 4 se referem, basicamente, aos usuários portadores de doenças oncológicas. Por conseguinte, o acompanhamento desses pacientes e a análise de seus gastos juntamente com as informações sobre a sua saúde é valiosa para encontrar possíveis relações entre tipos de câncer e quais podem ou não gerar internações e cirurgias.

O resumo dos agrupamentos consta no quadro 6.

Quadro 6 - Interpretação dos agrupamentos

AGRUPAMENTO	INSTÂNCIAS	% EM RELAÇÃO ÀS INSTÂNCIAS	% EM RELAÇÃO AO CUSTO	CARACTERÍSTICAS PRINCIPAIS
1 – Pacientes com condições simples de saúde	2929	96,57%	67,83%	"Gastos de até 5 mil reais e procedimentos relacionados à oftalmologia e à ortopedia, principalmente".
2 – Pacientes com altos custos hospitalares	71	2,34%	20,02%	"Gastos maiores de 40 mil reais com internações e cirurgias".
3 – Pacientes oncológicos com internações	15	0,49%	6,05%	"Pacientes oncológicos que tiveram internação".
4 – Pacientes oncológicos	18	0,59%	6,10%	"Tratamentos oncológicos".

Fonte: Autora (2023).

No quadro 7 podemos comparar algumas medidas estatísticas no que se relaciona à base processada e aos agrupamentos encontrados.

Quadro 7 – Estatística descritiva – Valores em reais (R$)

Grupo	Média	Moda	Mediana	Dispersão	Mínimo	Máximo
Base pre-processada	10.584,30	1.283,12	3.608,16	2.2217	1.001,27	248.265,09
Base processada	11.268,00	1.724,30	4210,90	1.9599	1.001,54	201.161,00
Agrupamento 1	7.915,63	1.724,30	4.004,06	1.3739	1.001,54	126.107,00
Agrupamento 2	96.285,90	41.231,30	87.191,80	0.40750	41.231,30	201.161,00
Agrupamento 3	137.900,00	64.860,10	136.518,00	0.26040	64.860,10	189.584,00
Agrupamento 4	115.899,00	67.643,90	107.227,00	0.29107	67.643,90	189.912,00

Fonte: Autora (2023).

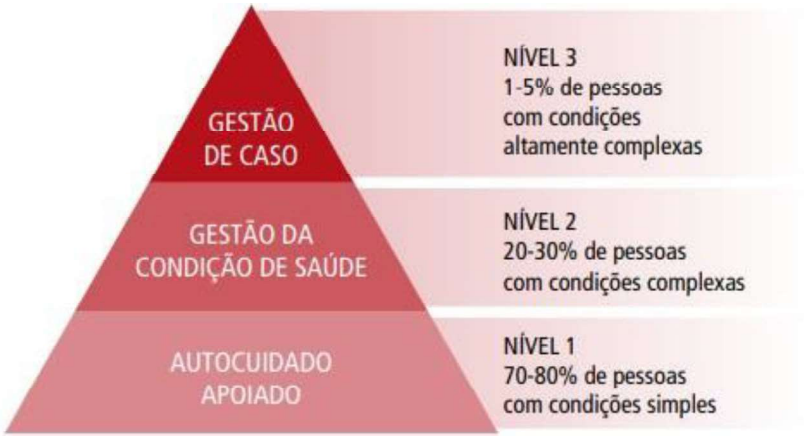
Uma maneira de se analisar os resultados encontrados é tentar interpretá- los de acordo com o modelo de pirâmide de risco – figura 54, muito utilizada em gestão de saúde populacional. Ela identifica três níveis de intervenções de acordo com a complexidade da condição crônica. A organização das pessoas usuárias, segundo as diferentes complexidades, permite orientar as intervenções em relação aos grupos de riscos e utilizar mais racionalmente os recursos humanos, concentrando-os nos grupos de maiores riscos (MENDES, 2011).

Sua lógica está em promover a saúde de toda a população, de estruturar as ações de autocuidado apoiado para os portadores de condições de saúde mais simples, de ofertar a gestão da condição de saúde para as pessoas que tenham uma condição estabelecida e de manejar os portadores de condições de saúde muito complexas por meio da tecnologia de gestão de caso (MENDES, 2011).

Assim, podemos caracterizar os agrupamentos 2, 3 e 4 no nível da 3 da pirâmide. Neste nível estão representados os usuários do sistema de saúde que necessitam de uma atenção maior dos gestores e profissionais da saúde. Podem, ainda, serem particularizados por necessitarem de cuidados relacionados a procedimentos de alto custo e de internações. Condições, estas, encontradas nestes agrupamentos.

O agrupamento 1, seguindo essa mesma metodologia, poderia ser dividido entre os níveis 2 e 3 da pirâmide - figura 54.

Figura 54 - Modelo de pirâmide de risco



Fonte: Bengoa, Porter e Kellogg (2008).

Por fim, cita-se, em relação aos programas de atenção à saúde, que a segregação encontrada, colabora com a assertividade deles, à medida que esclarece quais são as melhores atitudes a serem tomadas para melhorar a saúde da população de acordo com as peculiaridades de cada grupo, aumentando a chance de bons resultados nos programas e de engajamento dos pacientes nestas ações.

Um paciente engajado, seja participando de programas de atenção à saúde seja fazendo uso de oportunidades em atenção primária, tende a cuidar mais da sua saúde e consequentemente a usar o plano de saúde de maneira menos onerosa (PIRES et al., 2019).

7 CONCLUSÃO

O processo de descoberta de conhecimento em base de dados utilizando- se dados de planos de saúde foi realizado de forma satisfatória.

A pergunta de pesquisa foi respondida e assim, esse processo se mostrou como uma boa maneira de se encontrar informações acerca de como os usuários fazem uso desse sistema de saúde e de como podem ser agrupados e caracterizados.

Com a aplicação do algoritmo *K-means* sobre a base de dados pré- processada pode-se obter 4 agrupamentos. Esses agrupamentos trouxeram informações relevantes sobre a carteira de usuários do plano de saúde e podem servir de base para a tomada de decisões a respeito de ações a serem executadas quando se fala em gestão de saúde populacional.

De uma maneira resumida podemos caracterizar as pessoas agrupadas no agrupamento 1 naquelas que gastaram até R$5.000,00 anuais no plano de saúde. Neste grupo, que abarca a maioria dos usuários, há a possibilidade de destacar os usuários que fizeram uso de procedimentos relacionados à oftalmologia e à ortopedia. Já no agrupamento 2 observamos altos custos com internações hospitalares e reabilitação, além de custos relacionados à cardiologia.

No agrupamento 3 juntaram-se as pessoas que fazem tratamentos oncológicos e que tiveram custos com internações. No agrupamento 4 foram agrupadas as pessoas que fazem tratamentos com medicamentos quimioterápicos, mas que possuem baixos custos com internações e ao mesmo tempo ofendem, financeiramente, de forma relevante a carteira de plano de saúde.

A etapa de caracterização dos agrupamentos foi caracterizada como desafiadora. Conseguir extrair as informações dessas junções e por que eles foram organizados dessa forma exige amplo conhecimento acerca da base de dados e do conhecimento do negócio. Os gráficos *scatterplots* formados pelos agrupamentos e foram essenciais na descoberta de conhecimento. Usou-se também a ferramenta de distribuição dos dados. Conseguir concatenar as informações apresentadas com suas possíveis aplicações na área de saúde exige um conhecimento em mais de uma área do conhecimento como em estatística e em gestão populacional.

A divisão dos usuários do plano de saúde em grupos possibilita o conhecimento do perfil de utilização de cada paciente de acordo com as suas especificidades. Possibilita, ainda, a aproximação do gestor em saúde populacional com o paciente,

pois poderá direcionar ações e focar no grupo de pessoas que mais precisa de atenção. Por fim, essa técnica de agrupamento é importante dentro do que se chama de medicina personalizada. Assim, a questão de pesquisa deste estudo foi respondida de forma positiva, pois foi realizado o processo de descoberta de conhecimento utilizando os dados de procedimentos em saúde com sucesso.

Os agrupamentos formados permitiram a observação de novos conhecimentos a respeito do perfil de utilização da carteira de usuários. Somados a isso, permitiram o desenvolvimento de novas ideias de programas de atenção à saúde que podem ser aplicados aos usuários que fazem uso de consultas médicas de forma demasiada e que parecem não saber qual é o seu médico de referência, por exemplo. Programas como esse são desenvolvidos para melhorar a condição física e mental do usuário e contribuem para o cuidado e a geração de valor em saúde. Ainda, pode-se observar questões relacionadas ao custo dos usuários e onde estão sendo despendidos.

7.1 TRABALHOS FUTUROS

Trabalhos futuros podem ser realizados com essa mesma base de dados pré-processada e funções de agrupamento utilizando outros algoritmos e outras ferramentas de mineração de dados parecidas com o Orange como o *Weka* e o *Knime* ou, ainda, poder-se-á comparar os resultados obtidos neste trabalho a outros utilizando o *Python* ou o *R* que exigem conhecimento nessas linguagens de programação.

Cita-se, oportunidades, ainda, em modelos de aprendizado de máquina não supervisionado. Os códigos do agrupamento 1, ou seja, a referência aos usuários, devido às suas características e tamanho, pode ser utilizada em modelos de aprendizado em funções de agrupamento para dividir esse grupo, em funções de segmentação, para entender se quem utiliza procedimentos em oftalmologia, por exemplo, faz uso de outros procedimentos ao mesmo tempo.

Já as informações do agrupamento 2, podem ser usados para separar tipos de procedimentos em internações.

Trabalhos futuros, também podem ser realizados, mas usando modelos de aprendizado de máquina supervisionado, por exemplo. Sugere-se que os códigos do agrupamento 2 possam ser usados em modelo de predição para predizer risco de reinternações hospitalares. Neste agrupamento constam os maiores valores

referentes a procedimentos em pacientes internados ou que precisaram permanecer internados por motivos cirúrgicos.

Por fim, citam-se a possibilidade de utilizar mais de doze meses de dados em um novo modelo para análise de dados e de comparar essa base de dados que se relaciona ao período pré-COVID com o período pós-pandemia.

7.2 LIMITAÇÕES DA PESQUISA

Neste trabalho foi utilizado o conjunto de meses mais antigo disponível, ou seja, de janeiro a dezembro de 2019. Logo, como limitações da pesquisa cita- se a falta de dados anteriores a esse período. Assim, esse estudo poderá ser expandido com dados posteriores a 2019, somente.

REFERÊNCIAS

ABDELMAGID, A. S.; QAHMASH, A. I. M.. Utilizing the Educational Data Mining Techniques. Information Sciences Letters, [S.L.], v. 12, n. 3, p. 1415-1431, 1 mar. 2023. **Natural Sciences Publishing**. http://dx.doi.org/10.18576/isl/120330.

ALENCAR, Andre Luiz Siqueira. Atribuição de autoria por meio de ferramentas computacionais gratuitas – um estudo de caso. **Palimpsesto - Revista do Programa de Pós-Graduação em Letras da Uerj**, Universidade de Estado do Rio de Janeiro, v. 22, n. 41, p. 70-100, 11 abr. 2023..

ANSS. Agência Nacional de Saúde Suplementar. **NOTA TÉCNICA Nº 3/2019**. Disponível em: <https://www.sinlabpr.com.br/noticia/novidades/9-12-2019/-nota-tecnica-03-2019-%E2%80%93-agencia-nacional-de-saude-suplementar--ans-->>. Acesso em: 22 fev. 2022.

ARAÚJO, Flávio H.D.; SANTANA, André M.; SANTOS NETO, Pedro de A.. Using machine learning to support healthcare professionals in making preauthorisation decisions. **International Journal Of Medical Informatics**, [S.L.], v. 94, p. 1-7, out. 2016. Elsevier BV. http://dx.doi.org/10.1016/j.ijmedinf.2016.06.007.

BERTSIMAS, Dimitris; BJARNADÓTTIR, Margrét V.; KANE, Michael A.; KRYDER, J. Christian; PANDEV, Rudra; VEMPALA, Santos H.; WANG, Grant. Algorithmic Prediction of Health Care Costs. **Operations Research**, v. 56, n. 6, p. 1382-92, 2008.

BISHARA, Andrew; MAZE, Elijah H; MAZE, Mervyn. Considerations for the implementation of machine learning into acute care settings. **British Medical Bulletin**, [S.L.], v. 141, n. 1, p. 15-32, 28 jan. 2022. Oxford University Press (OUP). http://dx.doi.org/10.1093/bmb/ldac001.

BRUCE, Peter; BRUCE, Andrew. **Estatística Prática para Cientistas de Dados**. Rio de Janeiro: Alta Books, 2019.

CAMILO, Cássio Oliveira; SILVA, João Carlos da. **Mineração de Dados**: Conceitos, Tarefas, Métodos e Ferramentas. Goiás: Instituto de Informática Universidade Federal de Goiás, 2009.

DA COSTA, Susane Santos; CAZELLA, Silvio; RIGO, Sandro José. Minerando dados sobre o desempenho de alunos de cursos de educação permanente em modalidade EAD: um estudo de caso sobre evasão escolar na UNA- SUS. CINTED- Novas Tecnologias na Educação. **RENOTE**, [S.L.], v. 12, n. 2, 2014.

DOS SANTOS, Joana Raquel Raposo; DIAS, Carlos Matias; FILHO, Alexandre Chiavegatto. Machine Learning and national health data to improve evidence: finding segmentation in individuals without private insurance. **Health policy and technology.** 10. 79-86. 2021.

DOUPE, Patrick; FAGHMOUS, James; BASU, Sanjay. Machine Learning for Health

Services Researchers. **Value Health,** v. 22, n. 7, p. 808-815, 2019.

ESTATUTO DO IDOSO. Lei 10.741/2003.

ECKHARDT, Christina M.; MADJAROVA, Sophia J.; WILLIAMS, Riley J.; OLLIVIER, Mattheu; KARLSSON, Jón; PAREEK, Ayoosh; NWACHUKWU, Benedict U.. Unsupervised machine learning methods and emerging applications in healthcare. **Knee Surgery, Sports Traumatology, Arthroscopy**, [S.L.], v. 31, n. 2, p. 376-381, 15 nov. 2022. Springer Science and Business Media LLC. http://dx.doi.org/10.1007/s00167-022-07233-7.

FAYYAD, Usama; PIATETSKY-SHAPIRO, Gregory; SMYTH, Padhraic. From data mining to knowledge Discovery in databases. **AI Magazine**, v. 17, n. 3, p. 37-54, 1996.

FERNANDES, Anita Maria da Rocha; RADUENZ, Jean Carlo. Um levantamento sobre o uso de algoritmos de aprendizado de máquina em auditorias de planos de saúde. **Revista de Gestão em Sistemas de Saúde**, [S.L.], v. 9, n. 1, p. 119-131, 12 jun. 2020. University Nove de Julho. http://dx.doi.org/10.5585/rgss.v9i1.15296.

FERNANDES, Fernando Timoteo; CHIAVEGATTO FILHO, Alexandre Dias Porto. Perspectivas do uso de mineração de dados e aprendizado de máquina em saúde e segurança no trabalho. **Rev. Bras. Saúde Ocupacional**, v. 44, e:13, 2019.

FONSECA, Lais de Sá; OGATA, Alberto José Niituma. Proposta de modelo assistencial para uma operadora de saúde suplementar em expansão na cidade de São Paulo. **Revista de Administração em Saúde**, [S.L.], v. 21, n. 83, 6 jul. 2021. Associacao Brasileira de Medicina Preventia e Administração em Saúde - ABRAMPAS. http://dx.doi.org/10.23973/ras.83.291.

FORKAN, Abdur Rahim Mohammad; KHALIL, Ibrahim; KUMARAGE, Heshan. Patient clustering using dynamic partitioning on correlated and uncertain biomedical data. **Computer Methods And Programs In Biomedicine**, [S.L.], v. 190, p. 105483, jul. 2020. Elsevier BV. http://dx.doi.org/10.1016/j.cmpb.2020.105483.

GALVÃO, Noemi Dreyer; MARIN, Heimar de Fátima. Técnica de mineração de dados: uma revisão da literatura. **Acta Paul Enferm.**, v. 22, n. 5, p. 686-90, 2009.

GLOVER, Saundra; RIVERS, Patrick A.; ASOH, Derek A.; PIPER, Crystal N.; MURPH, Keva. Data mining for health executive decision support: an imperative with a daunting future! **Health Serv Manage Res.**, v. 23, n. 1, p. 42-46, fev. 2010.

HANDELMAN G. S.; KOK, H. K., CHANDRA, R. V.; RAZAVI, A. H.; LEE, M. J.; ASADI, H. eDoctor: machine learning and the future of medicine. **Journal of Internal Medicine**, v. 284, n. 6, p. 603-619, dez. 2018.

HASSANI-PAK, Keywan; RAWLINGS, Christopher. Knowledge Discovery in Biological Databases for Revealing Candidate Genes Linked to Complex Phenotypes. **Journal Of Integrative Bioinformatics**, [S.L.], v. 14, n. 1, p. 1-2, 1 mar. 2017. Walter de Gruyter GmbH. http://dx.doi.org/10.1515/jib-2016-0002.

ŀKRLJ, Blaž; KRALJ, Jan; LAVRAč, Nada. Embedding-based Silhouette community detection. **Machine Learning**, [S.L.], v. 109, n. 11, p. 2161-2193, 27 jul. 2020. Springer Science and Business Media LLC. http://dx.doi.org/10.1007/s10994-020-05882-8.

IHI – **Institute for healthcare improvement.** Disponível em: <http://www.ihi.org/.> Acesso em: 24 ago. 2020.

JOUDAKI, Hossein; RASHIDIAN, Arash; MINAEI-BIDGOLI, Behrouz; MAHMOODI, Mahmood; GERAILI, Bijan; NASIRI, Mahdi; ARAB, Mohammad. Using Data Mining to Detect Health Care Fraud and Abuse: A Review of Literature. **Global Journal of Health Science**, v. 7, n. 1, p.194-202, 2015.

KRITTANAWONG, Chayakrit; ZHANG, HongJu; WANG, Zhen; AYDAR, Mehmet; KITAI, Takeshi. Artificial Intelligence in Precision Cardiovascular Medicine. **Journal of the American College of Cardiology**, v. 69, n. 21, p. 2657- 2664, 2017.

LIM, Sunghoon; TUCKER, Conrad S.; KUMARA, Soundar. An unsupervised machine learning model for discovering latent infectious diseases using social media data. **Journal of Biomedical Informatics**, v. 66, p. 82-94, fev. 2017.

MARINS, Oudival Luiz Fraccaro de; BARROS, Everton Fernando; ROMÃO, Wesley; CONSTANTINO, Ademir Aparecido; SOUZA, Celso Lara. Aplicação de algoritmos de aprendizagem de máquina para mineração de dados sobre beneficiários de planos de saúde suplementar. **J. health inform**, v. 4, n. 2, abr.- jun. 2012.

MENDES, Eugênio Vilaça. **As redes de atenção à saúde.** Organização Pan-Americana de Saúde. 5499 p. 2011.

MENG, Lu; MESTRES, Nina; LU, Peng-Jun; SINGLETON, James; KRISS, Jennifer; ZHOU, Tianyi; WEISS, Débora; BLACK, Carla. Cluster analysis of adults unvaccinated for COVID-19 based on behavioral and social factors, National Immunization Survey-Adult COVID Module, United States. **Preventive Medicine**, [S.L.], v. 167, p. 107415, fev. 2023. Elsevier BV. http://dx.doi.org/10.1016/j.ypmed.2022.107415.

MORETTIN, Pedro A.; SINGER, Júlio M. **Introdução à Ciências de Dados.** Fundamentos e Aplicações. São Paulo: USP, 2020.

NETTO, Antônio Valério; BERTON, Lilian; TAKAHATA, Andre Kazuo. **Ciência de dados e a inteligência artificial na área da saúde.** 1. ed. São Paulo: Editora dos Editores, 2022.

PIRES, Denise Elvira Pires; VANDRESEN, Lara; MACHADO, Francele; MACHADO, Rosane Ramos; AMADIGI, Felipa Rafaela. **Gestão em saúde na atenção primária: o que é tratado na literatura.** Texto & Contexto Enfermagem v. 28: e20160426 2019.

PROVOST, Foster; FAWCETT, Tom. **Data science para negócios.** 1. ed. Rio de Janeiro: Alta Books, 2016.

QUINTERO, Yullys; ARDILA, Douglas; AGUILAR, Jose; CORTES, Santiago. Analysis of the socio economic impact due to COVID-19 using a deep clustering approach. **Applied Soft Computing,** v. 129. 2022.

RISTEVSKI, Blagoj; CHEN, Ming. Big Data Analytics in Medicine and Healthcare. **Journal of Integrative Bioinformatics**, v. 15, n. 3, 2018.

ROJAS, Eric; MUNOZ-GAMA, Jorge; SEPÚLVEDA, Marcos; CAPURRO, Daniel. Process mining in healthcare: A literature review. **Journal of Biomedical Informatics**, v. 61, p. 224-36, jun. 2016.

SANTANA, Roniel Venâncio Alencar; PONTES, Heráclito Lopes Jaguaribe. Applying K-means clustering to create product recommendation system based on purchase profiles. **Navus - Revista de Gestão e Tecnologia.** v. 10, p. 01-14. 2020.

SANTOS, Joana Raquel Raposo dos; DIAS, Carlos Matias; CHIAVEGATTO FILHO, Alexandre. Machine learning and national health data to improve evidence: finding segmentation in individuals without private insurance. **Health Policy And Technology**, [S.L.], v. 10, n. 1, p. 79-86, mar. 2021. Elsevier BV. http://dx.doi.org/10.1016/j.hlpt.2020.11.002.

SCHULTE, Timo; BOHNET-JOSCHKO, Sabine. How can Big Data Analytics Support People-Centred and Integrated Health Services: a scoping review. **International Journal Of Integrated Care**, [S.L.], v. 22, p. 23, 16 jun. 2022. Ubiquity Press, Ltd.. http://dx.doi.org/10.5334/ijic.5543.

SELEME, Ana Luísa Gonçalves Gomes Coelho; CUBAS, João Mário; CARVALHO, Deborah Ribeiro. SAÚDE MENTAL DO TRABALHADOR E O ALTO CUSTO DA ASSISTÊNCIA MÉDICA: uma análise por meio do aprendizado de máquina. **Revista Foco,** [S.L.], v. 16, n. 02, p. 920, 3 fev. 2023. South Florida Publishing LLC. http://dx.doi.org/10.54751/revistafoco.v16n2-059.

SETIAWAN, Karli Eka; KURNIAWAN, Afdhal; CHOWANDA, Andry; SUHARTONO, Derwin. Clustering models for hospitals in Jakarta using fuzzy c- means and k-means. **Procedia Computer Science**, [S.L.], v. 216, p. 356-363, 2023. Elsevier BV. http://dx.doi.org/10.1016/j.procs.2022.12.146.

SHARDA, Ramesh; DELEN, Dursun; TURBAN, Efraim. **Business intelligence e análise de dados para a gestão do negócio**. 4. ed. Porto Alegre: Bookman, 2019.

SOLEYMANI, Mohammad Haddad; YASERI, Mehdi; FARZADFAR, Farshad; MOHAMMADPOUR, Adel; SHARIFI, Farshad; KABIR, Mohammad Javad. Detecting medical prescriptions suspected of fraud using an unsupervised data mining algorithm. DARU. **Journal of Pharmaceutical Sciences,** v.26, p. 209- 214, 2018.

XIE, Yang; SCHREIER, Günter; HOY, Michael; LIU, Ying; NEUBAUER, Sandra; CHANG, David C.W.; REDMOND, Stephen J.; LOVELL, Nigel H.. Analyzing health insurance claims on different timescales to predict days in hospital. **Journal Of Biomedical Informatics**, [S.L.], v. 60, p. 187-196, abr. 2016. Elsevier BV.

http://dx.doi.org/10.1016/j.jbi.2016.01.002.

YUILL, Will; KUNZ, Holger. Using Machine Learning to Improve Personalised Prediction: a data-driven approach to segment and stratify populations for healthcare. **Studies In Health Technology And Informatics**, [S.L.], p. 0-0, 14 jan. 2022. IOS Press. http://dx.doi.org/10.3233/shti210851.

ZHANG, Hengwei; LI, Yan; MCCONNELL, William. Predicting potential palliative care beneficiaries for health plans: a generalized machine learning pipeline. **Journal Of Biomedical Informatics**, [S.L.], v. 123, p. 103922, nov. 2021. Elsevier BV. http://dx.doi.org/10.1016/j.jbi.2021.103922.

APÊNDICE A – CÓDIGOS SIP

As informações sobre a nomenclatura dos códigos SIP constam no quadro abaixo.

Essa nomenclatura foi unida aos códigos de procedimento principal (apêndice B) e juntas formaram os atributos do modelo de aprendizado de máquina apresentado neste estudo.

Embora o nome desse atributo seja “código” ele não possui número. Essa nomenclatura é originária da base de dados.

Quadro – Códigos SIP

Demais exames	Exames simples	Internações	Terapias
Consultas médicas em pronto socorro	Outros atendimentos ambulatoriais	Consultas médicas	-

APÊNDICE B – CÓDIGOS DE PROCEDIMENTO PRINCIPAL

As informações sobre a nomenclatura dos Códigos de procedimento principal SIP constam no quadro abaixo.

Essa nomenclatura foi unida aos códigos de procedimento principal (apêndice A) e juntas formaram os atributos do modelo de aprendizado de máquina apresentado neste estudo.

Embora o nome desse atributo seja “código” ele não possui número. Essa nomenclatura é originária da base de dados.

Quadro – Códigos de procedimento principal

anestesiologia	cardiologia	cirurgia cardiovascular	cirurgia da mão
cirurgia de cabeça e pescoço	cirurgia do aparelho digestivo	cirurgia geral	cirurgia oncológica
cirurgia pediátrica	cirurgia plástica	cirurgia torácica	cirurgia vascular
Citopatologia	clínica de cirurgia geral	clínica de cirurgia plástica	clínica de fisiatria
clínica de otorrinolaringologia	clínica medica	clínica medicina geral e família	coloproctologia
dermatologia	diagnóstico por imagem	ecografia	eletroencefalograma
endocrinologia	endocrinologia e metabologia	endoscopia	endoscopia / colonoscopia
espirometria	exames cardiológicos	fisioterapeuta geral e psicologia.	fonoaudiologia
gastroenterologia	geriatria	ginecologia	ginecologia e obstetrícia
hematologia	hematologia e hemoterapia	hemodinâmica	hepatologia
home care	homeopatia	hospital	hospital - clínicas psiquiátricas
infectologia	mastologia	medicina de família e comunidade	medicina do trabalho
medicina esportiva	medicina física e reabilitação-fisiatria	medicina intensiva	medicina nuclear

ANEXO A – GUIA DE PUBLICAÇÃO DE ARTIGOS PARA AUTORES DA REVISTA HEALTH POLICY AND TECHNOLOGY

O artigo redigido a partir desta dissertação será submetido à publicação para a revista *Health Policy and Technology*.

As orientações para os autores e os passos para encaminhamento do estudo constam no material abaixo.

Guide for Authors

Aims and scope

Published on behalf of the Fellowship of Postgraduate Medicine.

Health Policy and Technology is a peer-reviewed cross-disciplinary journal which focuses on past, present and future health policy and the role of technology in clinical and non-clinical health environments. HPT publishes relevant, timely and accessible articles and commentaries to support policy-makers, health professionals, health technology providers, patient groups and academia interested in health policy and technology.

The journal is owned by the registered charity the Fellowship of Postgraduate Medicine (FPM) established in 1918 with the aim of 'education medical professionals'.We invite papers on a range of policy and technology themes which may include:

- Cross-national comparisons on health policy using evidence-based approaches
- Country studies on health policy to determine the outcomes of technology-driven

initiatives
• Health technology, including drug discovery, diagnostics, medicines, devices, therapeutic delivery and eHealth systems
• Cross border eHealth including health tourism
• Health technology assessment (HTA) methods and tools for revaluating the effectiveness of clinical and non clinical health technologies
• eHealth systems
• Regulation and health economics

Authors for whom English is a second language may choose to have their manuscript professionally edited before submission or during the review process. Authors wishing to pursue a professional English-language editing service should make contact and arrange payment with the editing service of their choice. For more details regarding the recommended services, please refer to http://support.elsevier.com/

I. ETHICS IN PUBLISHING

For information on Ethics in Publishing and Ethical guidelines for journal publication see https://www.elsevier.com/publishingethics and https://www.elsevier.com/ethicalguidelines.

II. AUTHORSHIP AND ACKNOWLEDGEMENTS

All authors should have made substantial contributions to all of the following: (1) the acquisition of data, or analysis and interpretation of data, (2) drafting the article or revising it critically for important intellectual content, (3) final approval of the version to be submitted.

Acknowledgements

All contributors who do not meet the criteria for authorship as defined above should be listed in an acknowledgements section. Examples of those who might be acknowledged include a person who provided purely technical help, writing assistance, or a department chair who provided only general support. Authors should disclose whether they had any writing assistance and identify the entity that paid for this assistance.

When submitting a paper authors must complete the Authorship form download from here. This form confirms that all authors agree to publication if the paper is accepted and allows authors to declare any conflicts of interest, sources of funding and ethical approval (if required). Please download the form and submit it with your paper. Submissions that do not include a completed form will be returned without review.

Declaration of generative AI in scientific writing

The below guidance only refers to the writing process, and not to the use of AI tools to analyse and draw insights from data as part of the research process.

Where authors use generative artificial intelligence (AI) and AI-assisted technologies in the writing process, authors should only use these technologies to improve readability and language. Applying the technology should be done with human oversight and control, and authors should carefully review and edit the result, as AI can generate authoritative-sounding output that can be incorrect, incomplete or biased. AI and AI- assisted technologies should not be listed as an author or co-author, or be cited as an author.

Authorship implies responsibilities and tasks that can only be attributed to and performed by humans, as outlined in Elsevier's AI policy for authors.

Authors should disclose in their manuscript the use of AI and AI-assisted technologies in the writing process by following the instructions below. A statement will appear in the published work. Please note that authors are ultimately responsible and accountable for the contents of the work.

Disclosure instructions

Authors must disclose the use of generative AI and AI-assisted technologies in the writing process by adding a statement at the end of their manuscript in the core manuscript file, before the References list. The statement should be placed in a new section entitled 'Declaration of Generative AI and AI-assisted technologies in the writing process'. *Statement: During the preparation of this work the author(s) used [NAME TOOL / SERVICE] in order to [REASON]. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.*

This declaration does not apply to the use of basic tools for checking grammar, spelling, references etc. If there is nothing to disclose, there is no need to add a statement.

III. SUBMISSION DECLARATION

Submission of an article implies that the work described has not been published previously (except in the form of an abstract or as part of a published lecture or academic thesis), that it is not under consideration for publication elsewhere, that its publication is approved by all authors and tacitly or explicitly by the responsible authorities where the work was carried out, and that, if accepted, it will not be published elsewhere including electronically in the same form, in English or in any other language, without the written consent of the copyright-holder.

IV. RETAINED AUTHOR RIGHTS

As an author you (or your employer or institutions) retain certain rights; for details you are referred to: https://www.elsevier.com/authorsrights.

V. FUNDING BODY AGREEMENTS AND POLICIES

Elsevier has established agreements and developed policies to allow authors whose articles appear in journals published by Elsevier, to comply with potential manuscript archiving requirements as specified as conditions of their grant awards. To learn more about existing agreements and policies please visit https://www.elsevier.com/fundingbodies.

VI. MANUSCRIPT SUBMISSION AND SPECIFICATIONS

To submit a manuscript to Health Policy and Technology, please go to: https://www.editorialmanager.com/hlpt/default.aspx

If submissions are larger than 500 KB, they should be compressed using PKZIP or WINZIP.

Copyright: Upon acceptance of an article, authors will be asked to transfer copyright (for more information on copyright see http://authors.elsevier.com). This transfer will ensure the widest possible dissemination of information. A letter will be sent to the corresponding author confirming receipt of the manuscript. A form facilitating transfer of copyright will be provided.

If excerpts from other copyrighted works are included, the author(s) must obtain written permission from the copyright owners and credit the source(s) in the article. Elsevier has preprinted forms for use by Authors in these cases: contact Elsevier's Rights Department, Oxford UK: e-mail: permissions@elsevier.com. Requests may also be completed on-line via the Elsevier homepage (https://www.elsevier.com/locate/permissions).

Elsevier supports responsible sharing
Find out how you can share your research published in Elsevier journals.

Peer Review
This journal operates a double-blind review process. All contributions will be initially assessed by the editor for suitability for the journal. Papers deemed suitable are then typically sent to a minimum of two independent expert reviewers to assess the scientific quality of the paper. The Editor-in-Chief is responsible for the final decision regarding acceptance or rejection of articles. The Editor-in-Chief's decision is final. Editors are not involved in decisions about papers which they written themselves or have been written by family members or colleagues or which relate to products or services in which the editor has an interest. Any such submission is subject to all of the journal's usual procedures, with peer review handled independently of the relevant editor and their research groups.

Each Submission should contain separate documents as follows:

1) COVER LETTER
The cover letter should include: 1) title of the manuscript; 2) name of the document file(s) containing the manuscript and the software (and version) used; 3) name and all contact information for the corresponding author and a statement as to whether the data, models, or methodology used in the research are proprietary; 4) names of all sponsors of the research and a statement of all direct or indirect financial relationships the authors have with the sponsors; and 5) if applicable, a statement that the publication of study results was not contingent on the sponsor's approval or censorship of the manuscript.

2) TITLE PAGE
The title page should contain the following:
1) Title
2) Full names (first and surname) of all authors including academic degrees and affiliation(s)
3) Name, mailing and email addresses, telephone and fax numbers of corresponding author (with whom all correspondence will take place unless other arrangements are made)
4) All sources of financial or other support for the manuscript (if no funding was received, this should be noted on the title page)
5) Competing interests
6) Ethical approval
7) Acknowledgements

8) At least four key words for indexing purposes; and
9) A running title of not more than 45 characters including spaces.

If there are no declarations to make, the following statements should be inserted into the manuscript:
Funding: None
Competing interests: None declared
Ethical approval: Not required

3) MANUSCRIPTS
Manuscripts must be written in English, typed in either Microsoft Word (Version 5.0 or later) or WordPerfect (version 5.1 or later). Manuscripts should be double- spaced with 1-inch margins on all sides and size 10 font (Arial or Times New Roman fonts are preferred). Minimal formatting should be used, i.e., no justification, italics, bold, indenting, etc. There should be no hard returns at the end of lines. Double- spacing after each element is requested (e.g., headings, titles, paragraphs, legends).

Editorials: Editorials are generally invited by the Editorial Team. They are 1,500 words in length with no abstract or keywords.
Original research articles: Original research articles have a limit of 4,500 words and no more than 40 references.
Review articles: Review articles have a limit of 3,500 words with an unlimited number of references.
Research Letters: Research Letters have a limit of 1,000 words, no more than 1 Figure or Table, and no more than 10 references.
Commentary articles: Commentary articles have a maximum word count of 1500 words and no abstract is required. The maximum number of references is 10.

ABSTRACT
An abstract of 250 words or less is required, summarizing the work reported in the manuscript. Original research manuscripts should use a structured format for the abstract, i.e., Objectives, Methods, Results, Conclusions.

LAY SUMMARIES
You are required to provide your abstract in two parts:

Firstly: a scientific abstract of 250 words or less is required, summarizing the work reported in the manuscript. Original research manuscripts should use a structured format for the abstract, i.e., Objectives, Methods, Results, Conclusions.

Secondly a public interest abstract of 150 words or less summarising the main message of the article expressed in plain English to describe your findings to a non-medical audience. This should sit at the end of your abstract text after a heading Public Interest Summary.

See link if you wish advice on how to write a public interest (lay) summary. https://www.elsevier.com/connect/authors-update/in-a-nutshell-how-to-write- a-lay-summary

TEXT

The body of the manuscript should be divided into sections that facilitate reading and comprehension of the material. This should normally include sections with the major headings: Introduction, Methods, Results, Conclusions, Acknowledgments (if needed), and References. There should be no footnotes. Figures (inclusive of figure legends) and Tables must be submitted each as separate documents.

REFERENCES

The format of references should be that of the Vancouver guidelines. Include: The names of all the authors when six or fewer, followed by their initials. Otherwise list only the first three and add *et al*
The title of the article or chapter
The journal name abbreviated as in *Index Medicus,* the year and volume, and the first and last pages
For a book, the names of any editors (as for authors), the city and name of the publisher, and the year and pages

Examples for an article in a journal (1) or book (2) or for a book (3) would be:
1. Jiang FN, Liu DJ, Neyndorff H, Chester M, Jiang S-Y, Luy JG. Photodynamic killing of human squamous cell carcinoma cells using a monoclonal antibody-photosensitizer conjugate. *J Natl Cancer Inst* 1991;83:1218-25.
2. Gullick WJ, Venter DJ. The c-erbB2 and its expression in human tumours. In Waxman J, Sikora K, editors. *The molecular biology of cancer. Oxford:* Blackwell Scientific Publications; 1989: p.38-53.
3. Lumley JSP, Green CJ, Lear P, Angell-James JE, *Essentials of Experimental Surgery.* London: Butterworths; 1990.

Data references This journal encourages you to cite underlying or relevant datasets in your manuscript by citing them in your text and including a data reference in your Reference List. Data references should include the following elements: author name(s), dataset title, data repository, version (where available), year, and global persistent identifier. Add [dataset] immediately before the reference so we can properly identify it as a data reference. This identifier will not appear in your published article.

[dataset] [5] Oguro M, Imahiro S, Saito S, Nakashizuka T. Mortality data for Japanese oak wilt disease and surrounding forest compositions, Mendeley Data, v1; 2015. http://dx.doi.org/10.17632/xwj98nb39r.1.

DATA, MODELS, AND METHODOLOGYM

All authors must agree to make their data available at the Editor's request for examination and re-analysis by referees or other persons designated by the Editor. All models and methodologies must be presented in sufficient detail to be fully comprehensible to readers.

Figure Captions, Tables, Figures and Schemes:

Preparation of Electronic Illustrations
. o Make sure you use uniform lettering and sizing of your original artwork.
o Save text in illustrations as "graphics" or enclose the font.
o Only use the following fonts in your illustrations: Arial, Courier, Helvetica, Times, Symbol.
o Number the illustrations according to their sequence in the text.
o Use a logical naming convention for your artwork files.

o Provide all illustrations as separate files and as hardcopy printouts on separate sheets.
o Provide captions to illustrations separately.
o Produce images near to the desired size of the printed version.

A detailed guide on electronic artwork is available on our website: https://www.elsevier.com/artwork. You are urged to visit this site; some excerpts from the detailed information are given here.

Formats. Regardless of the application used, when your electronic artwork is finalised, please "save as" or convert the images to one of the following formats (Note the resolution requirements for line drawings, halftones, and line/halftone combinations given below.):
EPS: Vector drawings. Embed the font or save the text as "graphics". TIFF: Colour or greyscale photographs (halftones): always use a minimum of 300 dpi. TIFF: Bitmapped line drawings: use a minimum of 1000 dpi. TIFF: Combinations bitmapped line/half-tone (colour or greyscale): a minimum of 500 dpi is required.
JPEG, DOC, XLS or PPT: If your electronic artwork is created in any of these Microsoft Office applications please supply "as is".

Please do not:
• Supply embedded graphics in your word processor (spreadsheet, presentation) document;
• Supply files that are optimised for screen use (like GIF, BMP, PICT, WPG); the resolution is too low;
• Supply files that are too low in resolution;
• Submit graphics that are disproportionately large for the content.

If, together with your accepted article, you submit usable colour figures then Elsevier will ensure, at no additional charge that these figures will appear in colour on the Web (e.g., ScienceDirect and other sites).

Captions.
Ensure that each illustration has a caption. Supply captions separately, not attached to the figure. A caption should comprise a brief title (not on the figure itself) and a description of the illustration. Keep text in the illustrations themselves to a minimum but explain all symbols and abbreviations used.

Line Drawings.
The lettering and symbols, as well as other details, should have proportionate dimensions, so as not to become illegible or unclear after possible reduction; in general, the figures should be designed for a reduction factor of two to three. The degree of reduction will be determined by the Publisher. Illustrations will not be enlarged. Consider the page format of the journal when designing the illustrations. Do not use any type of shading on computer-generated illustrations.

Photographs (halftones).
Remove non-essential areas of a photograph. Do not mount photographs unless they form part of a composite figure. Where necessary, insert a scale bar in the illustration (not below it), as opposed to giving a magnification factor in the caption. Note that photocopies of photographs are not acceptable.

Revised Manuscripts:
Authors who have been asked to revise their manuscript by the Editors should submit a file which clearly shows the changes that have been made via the 'track changes' function or text highlighting, and a file containing a clean copy of the manuscript.

Preparation of Supplementary Data: Elsevier accepts electronic supplementary material to support and enhance your scientific research. Supplementary files offer the author additional possibilities to publish supporting applications, movies, animation sequences, high-resolution images, background datasets, sound clips and more. Supplementary files supplied will be published online alongside the electronic version of your article in Elsevier Web products, including ScienceDirect: http://www.sciencedirect.com. In order to ensure that your submitted material is directly usable, please ensure that data is provided in one of our recommended file formats. Authors should submit the material in electronic format together with the article and supply a concise and descriptive caption for each file. For more detailed instructions please visit our artwork instruction pages at the Author Gateway at https://www.elsevier.com/artwork.

RESEARCH DATA
This journal encourages and enables you to share data that supports your research publication where appropriate, and enables you to interlink the data with your published articles. Research data refers to the results of observations or experimentation that validate research findings. To facilitate reproducibility and data reuse, this journal also encourages you to share your software, code, models, algorithms, protocols, methods and other useful materials related to the project. Below are a number of ways in which you can associate data with your article or make a statement about the availability of your data when submitting your manuscript. If you are sharing data in one of these ways, you are encouraged to cite the data in your manuscript and reference list. Please refer to the "References" section for more information about data citation. For more information on depositing, sharing and using research data and other relevant research materials, visit the research data page.

Data linking
If you have made your research data available in a data repository, you can link your article directly to the dataset. Elsevier collaborates with a number of repositories to link articles on ScienceDirect with relevant repositories, giving readers access to underlying data that gives them a better understanding of the research described. There are different ways to link your datasets to your article. When available, you can directly link your dataset to your article by providing the relevant information in the submission system. For more information, visit the database linking page .For supported data repositories a repository banner will automatically appear next to your published article on ScienceDirect.In addition, you can link to relevant data or entities through identifiers within the text of your manuscript, using the following format: Database: xxxx (e.g., TAIR: AT1G01020; CCDC: 734053; PDB: 1XFN).

Mendeley Data
This journal supports Mendeley Data, enabling you to deposit any research data (including raw and processed data, video, code, software, algorithms, protocols, and methods) associated with your manuscript in a free-to-use, open access repository. During the submission process, after uploading your manuscript, you will have the opportunity to upload your relevant datasets directly to *Mendeley Data*. The datasets will be listed and

directly accessible to readers next to your published article online.
For more information, visit the Mendeley Data for journals page.

Data statement
To foster transparency, we encourage you to state the availability of your data in your submission. This may be a requirement of your funding body or institution. If your data is unavailable to access or unsuitable to post, you will have the opportunity to indicate why during the submission process, for example by stating that the research data is confidential. The statement will appear with your published article on ScienceDirect. For more information, visit the Data statement page.

Special Subject Repositories: Elsevier has established agreements and developed policies to allow authors who publish in Elsevier journals to comply with potential manuscript archiving requirements as specified as conditions of their grant awards. To learn more about existing agreements and policies please visit https://www.elsevier.com/fundingbodies.

Highlights
Highlights are mandatory for this journal. They consist of a short collection of bullet points that convey the core findings of the article and should be submitted in a separate file in the online submission system. Please use 'Highlights' in the file name and include 3 to 5 bullet points (maximum 85 characters, including spaces, per bullet point). See https://www.elsevier.com/highlights for examples.**Sponsored Articles:**

Health Policy and Technology offers authors the option to sponsor non-subscriber access to individual articles. The access sponsorship contribution fee per article is
$3,000. This contribution is necessary to offset publishing costs - from managing article submission and peer review, to typesetting, tagging and indexing of articles, hosting articles on dedicated servers, supporting sales and marketing costs to ensure global dissemination via ScienceDirect, and permanently preserving the published journal article. The sponsorship fee excludes taxes and other potential author fees such as colour charges which are additional.

Authors can specify that they would like to select this option after receiving notification that their article has been accepted for publication, but not before. This eliminates a potential conflict of interest by ensuring that the journal does not have a financial incentive to accept an article for publication.

Proofs:
One set of page proofs in PDF format will be sent by e-mail to the corresponding author (if we do not have an e-mail address then paper proofs will be sent by post). Elsevier now sends PDF proofs which can be annotated; for this you will need to download Adobe Reader version 7 availablefree from http://www.adobe.com/products/acrobat/readstep2.html. Instructions on how to annotate PDF files will accompany the proofs. The exact system requirements are given at the Adobe site: http://www.adobe.com/products/acrobat/acrrsystemreqs.html#70win.

If you do not wish to use the PDF annotations function, you may list the corrections (including replies to the Query Form) and return to Elsevier in an e-mail. Please list your corrections quoting line number. If, for any reason, this is not possible, then mark the

corrections and any other comments (including replies to the Query Form) on a printout of your proof and return by fax, or scan the pages and e-mail, or by post.

Please use this proof only for checking the typesetting, editing, completeness and correctness of the text, tables and figures. Significant changes to the article as accepted for publication will only be considered at this stage with permission from the Editor. We will do everything possible to get your article published quickly and accurately. Therefore, it is important to ensure that all of your corrections are sent back to us in one communication: please check carefully before replying, as inclusion of any subsequent corrections cannot be guaranteed. Proofreading is solely your responsibility. Note that Elsevier may proceed with the publication of your article if no response is received.

Offprints:
The corresponding author, at no cost, will be provided with a PDF file of the article via e-mail. The PDF file is a watermarked version of the published article and includes a cover sheet with the journal cover image and a disclaimer outlining the terms and conditions of use. Additional paper offprints can be ordered by the authors. An order form with prices will be sent to the corresponding author.

Appeals:
If you believe that the editorial decision about your manuscript was based on factual errors, you can contact us at HLPT@elsevier.com. Please state the manuscript number and describe in detail why you believe the decision was erroneous. Your appeal will be seen again by the handling editor, who will revisit the previous decision in light of your comments. Please note that we do not allow multiple appeals: a second decision will be final.

Reporting sex- and gender-based analyses Reporting guidance
For research involving or pertaining to humans, animals or eukaryotic cells, investigators should integrate sex and gender-based analyses (SGBA) into their research design according to funder/sponsor requirements and best practices within a field. Authors should address the sex and/or gender dimensions of their research in their article. In cases where they cannot, they should discuss this as a limitation to their research's generalizability. Importantly, authors should explicitly state what definitions of sex and/or gender they are applying to enhance the precision, rigor and reproducibility of their research and to avoid ambiguity or conflation of terms and the constructs to which they refer (see Definitions section below). Authors can refer to the Sex and Gender Equity in Research (SAGER) guidelines and the the SAGER guidelines checklist. These offer systematic approaches to the use and editorial review of sex and gender information in study design, data analysis, outcome reporting and research interpretation - however, please note there is no single, universally agreed-upon set of guidelines for defining sex and gender.

Definitions
Sex generally refers to a set of biological attributes that are associated with physical and physiological features (e.g., chromosomal genotype, hormonal levels, internal and external anatomy). A binary sex categorization (male/female) is usually designated at birth ("sex assigned at birth"), most often based solely on the visible external anatomy of a newborn. Gender generally refers to socially constructed roles, behaviors, and identities of women, men and gender-diverse people that occur in a historical and cultural context and may vary across societies and over time. Gender influences how people view themselves and each other, how they behave and interact and how power is distributed in society. Sex and gender are often incorrectly portrayed as binary (female/male or woman/man) and

unchanging whereas these constructs actually exist along a spectrum and include additional sex categorizations and gender identities such as people who are intersex/have differences of sex development (DSD) or identify as non-binary. Moreover, the terms "sex" and "gender" can be ambiguous?thus it is important for authors to define the manner in which they are used. In addition to this definition guidance and the SAGER guidelines, the resources on this page offer further insight around sex and gender in research studies.