DESCOBERTA DE CONHECIMENTO APLICADA À BASE DE DADOS ABERTOS DA ANVISA SOBRE PREÇOS DE MEDICAMENTOS POR MEIO DE ANÁLISE DE REDES DE INFORMAÇÃO

Discovery Of Knowledge Applied To The Open Database Of Anvisa About Drug Prices Through Information Network Analysis

Lucas Vale[1], Henrique Monteiro Cristovão[2]
(1) Universidade Federal do Espírito Santo, Campus de Goiabeiras, lucas.s.vale@edu.ufes.br
(2) Universidade Federal do Espírito Santo, Campus de Goiabeiras, henrique.cristovao@ufes.br

Resumo:
Propõe-se nesta pesquisa investigar e revelar possíveis relações entre variáveis da base de dados abertos da Anvisa sobre preço de medicamentos. A pesquisa tem abordagem qualitativa e natureza aplicada. A base de dados analisada é mantida e alimentada pela ANVISA, disponível no portal de dados abertos do governo, sendo composta por 26310 registros correspondentes ao período de 2017 a 2021. Fundamentado na metodologia de descoberta de conhecimento de Fayyad onde, nas fases pré-processamento e transformação, usou-se o software OpenRefine, e nas fases mineração de dados, interpretação e avaliação aplicou-se a análise de redes complexas com suporte dos softwares OpenRefine e Microsoft Power BI Desktop. Observa-se, apesar da ausência de alguns registros, que os medicamentos mais produzidos pelos laboratórios registrados são os de tarja vermelha, onde estão medicamentos que tratam condições de alta prevalência no Brasil, como hipertensão e diabetes, enquanto que os medicamentos de tarja preta têm produção e demanda muito mais limitadas pela suas restrições de uso. Destaca-se a alta produção de classes terapêuticas em que o custo de produção gira em torno de R$100, indicando que a maioria dos laboratórios têm em seu modelo de negócios públicos-alvo das classes C e D, ao passo que apenas um laboratório é responsável pela produção de todos os medicamentos em que o custo de produção ultrapassa 1 milhão de reais. Entende-se ainda que é preciso dedicar mais esforços na análise da base de dados a fim de potencializar a descoberta de novas relações entre variáveis.

Palavras-chave: *Descoberta de conhecimento; Anvisa; Análise de redes complexas; Ciência de dados;*

Abstract:
The research proposes to investigate and reveal possible relationships between variables in Anvisa's open database on drug prices. The research has a qualitative approach and an applied nature. The analyzed database is maintained and fed by ANVISA, available on the government's open data portal, consisting of 26,310 records corresponding to the period from 2017 to 2021. Based on Fayyad's knowledge discovery methodology where, in the pre-processing and transformation, the OpenRefine software was used, and in the data mining, interpretation and evaluation phases, the analysis of complex networks was applied with support from the OpenRefine and Microsoft Power BI Desktop software. It is observed, despite the absence of some records, that the drugs most produced by registered laboratories are those with the red stripe, which contain drugs that treat highly prevalent conditions in Brazil, such as hypertension and diabetes, while the black stripe drugs have production and demand much more limited by their use restrictions. The high production of therapeutic classes stands out, in which the production cost is around R$100, indicating that most laboratories have in their business model target audiences from classes C and D, while only one laboratory is responsible for the production of all medicines in which the production cost exceeds 1 million reais. It is also understood that it is necessary to dedicate more efforts to the analysis of the database in order to enhance the discovery of new relationships between variables.

Keywords: *Discovery of knowledge; Anvisa; Complex networks analysis; Data Science;*

1 Introdução

É concomitante ao desenvolvimento da civilização o acúmulo de dados onde, ter acesso à informação desejada de maneira precisa é um diferencial que pode fazer prosperar negócios em áreas muito distintas. Isso faz dos profissionais responsáveis pelo processo de tomada de decisão dependentes de ferramentas assertivas para os procedimentos de coleta, organização, processamento e utilização da informação. (COSTA et al., 2009)

Temas relacionados à saúde e ao bem-estar são sempre sensíveis e relevantes para

a sociedade. No contexto onde a população está cada vez mais envelhecida, entende-se que é primária a preocupação com um Sistema Nacional de Saúde eficiente e robusto (BÁRRIOS et al. 2020). No Brasil, o Sistema Único de Saúde (SUS) tem integrado a si a autarquia ANVISA (Agência Nacional de Vigilância Sanitária), responsável por, entre outras coisas, a gestão do controle de qualidade e da vigência de preços de medicamentos.

O processo de descoberta de conhecimentos (KDD - *Knowledge Discovery in Databas*e) é um ponto de partida metodológico para a identificação de padrões em um determinado conjunto de dados em que exista potencial de compreensão e utilidade para o usuário. No processo sugerido por Fayyad et al. (1996), existem 6 etapas: seleção de dados, pré-processamento, transformação, mineração de dados, interpretação e avaliação e implantação do conhecimento.

A etapa de pré-processamento é onde os dados são manipulados com o propósito de solucionar problemas, corrigir erros, modificar a estrutura, adequar tipos entre outros, preparando-os para que a fase de descoberta de conhecimento seja mais eficiente. Nesta etapa pode-se aprender mais sobre os dados em uso, por meio de visualizações. De modo geral, a fase de pré-processamento é tida como semi-automática, isto é, dependente da aptidão do operador em identificar problemas no conjunto de dados e a sua natureza, como expõe Batista (2003).

Na etapa da transformação podem ocorrer alterações significativas sobre os dados e, principalmente, a criação de novos subconjuntos de dados para atender a requisitos específicos de determinados softwares de mineração.

A etapa de mineração de dados tem grande diversidade de métodos disponíveis para uso. Na presente pesquisa há ênfase na técnica de análise de redes complexas por meio de, principalmente, o uso de métricas, comunidades, projeções bipartidas e inspeção visual. Uma rede complexa, segundo Barabási (2003), é um conjunto de nós interligados por arestas que constitui uma estrutura topográfica sofisticada. O estudo das redes neste formato teve início com a solução do problema das pontes de Königsberg por Euler, em 1735, que derivou a teoria dos grafos (METZ, 2007). A análise de redes complexas, portanto, se destina a compreender a relação entre nós e as suas consequências, pouco se preocupando com as características particulares dos nós, e sim com a sua totalidade estrutural, tomando emprestados fundamentos da análise de redes sociais, como sugerem Wasserman et al. (1994). Nesta etapa, os dados dispostos em nós e arestas fornecem possibilidades de inspeção visual, desde que adequadamente formatados sobre a topologia e as relações existentes entre os nós.

A inspeção visual, ou visualização dos dados, é estrutural na análise de redes de informação para tornar padrões despercebidos evidentes (CHEN, 2013). O processo de visualização pode, além do mais, ser constituinte da fase de pré processamento ou de transformação, uma vez que os limites entre as fases delimitadas pelo KDD não são bem estabelecidos, segundo HAND et al. (2001), sendo de toda forma um artifício da mineração sobretudo da interpretação dos dados.

Imagina-se que o conjunto de dados escolhido possa, ao ser analisado, revelar padrões ou tendências não evidentes ou despercebidas que, em função da sua forte relevância social, podem trazer algum ganho para os indivíduos diretamente envolvidos com o preço de medicamentos no Brasil.

2 Objetivos

Investigar e revelar relações entre variáveis da base de dados abertos da ANVISA sobre preços de medicamentos.

3 Procedimentos Metodológicos

A pesquisa tem abordagem qualitativa e natureza aplicada. A base de dados analisada é mantida e alimentada pela ANVISA, disponível no portal de dados abertos do governo[1] de maneira estruturada e é composta por 26310 registros correspondentes ao período de 2017 a 2021, e com tamanho aproximado de 10Mb.

[1] Disponível em: *https://dados.gov.br/dataset/preco-de-medicamentos-no-brasil-consumidor*

Após a obtenção dos dados as etapas do KDD foram executadas conforme descritas nas próximas subseções e ilustradas na Figura 1.

3.1 Pré-processamento

Nesta etapa foram removidos ruídos da base de dados e selecionados os atributos mais relevantes para a análise, bem como a formatação adequada dos dados.

Fazendo uso da ferramenta OpenRefine[2], foram excluídas as colunas CNPJ, Registro, EAN, PF 0%, PF 12%, PF 17,5%, PF 17,5% ALC, PF 18%, PF 18% ALC, PF 20%, PMC 12%, PMC 17,5%, PMC 17,5% ALC, PMC 18%, PMC 18% ALC, PMC 20%, Código GGREM, por serem irrelevantes para o estudo, do ponto de vista dos autores e objetivos da pesquisa. O nome das colunas foi alterado para o padrão camelCase[3] para melhor integração entre as ferramentas de análise utilizadas. As colunas contendo valores numéricos foram convertidas para a classe de números (variável quantitativa contínua) e os valores coluna de *tarja,* que continham irregularidades derivadas da inserção dos registros foram padronizados em 4 categorias: *Tarja Preta, Tarja Vermelha, Venda Livre e Sem Tarja*.

3.2 Transformação

Neste ponto, alguns dados são transcritos para formatos mais adequados para a análise, considerando-se peculiaridades do método de análise de redes de informação. Utilizando a linguagem GREL[4], os valores numéricos de preço (quantitativos contínuos) foram transformados em categorias de preço (qualitativos ordinais) seguindo a seguinte organização para as categorias:

- Abaixo de R$100
- Entre R$100 e R$1000
- Entre R$1000 e R$50000
- Entre R$50000 e 1 milhão
- Acima de 1 milhão de reais

Em seguida, foram geradas redes de informação por meio de mapeamento realizado pelo software OpenRefine para o formato de rede GML[5] reconhecido pelo software Gephi[6]. Como exemplo, consta na Figura 2 do Apêndice A o mapeamento realizado para a criação da rede tripartida citada na seção 3.3.

3.3 Mineração de dados, interpretação e avaliação

Na fase de mineração utilizou-se a metodologia de análise de redes complexas que, no caso da presente pesquisa, foram redes de informação. Também foi criado um *dashboard* para a visualização sintetizada de alguns dados.

Os objetivos da análise de redes resumiram-se a analisar a relação entre laboratórios (de produção dos medicamentos) e tarja, a relação entre classes terapêuticas, faixa de preço de revenda e a relação entre laboratórios e faixa de preço de fábrica, classe terapêutica e regime de preço via projeção bipartida com geração de uma rede monopartida de classe terapêutica.

As redes brutas geradas foram organizadas utilizando os métodos de distribuição Yifan Hu e Fruchterman Reingold, cujos algoritmos estão disponíveis no software Gephi. A visualização dos dados utilizando *dashboard* foi desenvolvida no software Microsoft Power BI Desktop[7] a partir

[2] O software OpenRefine é uma ferramenta de código aberto utilizada para a limpeza e transformação de dados. Disponível em https://openrefine.org/.

[3] O formato camelCase integra o uso de vários softwares e linguagens por meio da padronização sugerida na terminologia de nomes de variáveis. Disponível em: https://pt.wikipedia.org/wiki/CamelCase.

[4] *General Refine Expression Language* (GREL). É uma linguagem de script similar ao javascript utilizada no ambiente do OpenRefine.

[5] GML (Graph Modelling Language) é um formato para representação de grafos de fácil leitura por humanos e com uma capacidade semântica razoável para configurar as características da rede, dos nós e das arestas. Disponível em: https://en.wikipedia.org/wiki/Graph_Modelling_Language/.

[6] GEPHI é um software de código aberto utilizado para visualização, análise e manipulação de redes e grafos. Disponível em https://gephi.org/.

[7] O Microsoft Power BI Desktop é um conjunto de serviços de software de uso gratuito destinados à

da base de dados em extensão '.xlsx' exportada pela ferramenta OpenRefine, após as etapas de pré-processamento e transformação. O *dashboard* foi construído a partir da seleção de laboratórios, tarja e regime de preço, exibindo a comparação entre preço de fábrica e preço de revenda para cada produto, a distribuição de todos produtos em faixas de preço de fábrica e o número de medicamentos por taja.

4 Resultados

Na etapa de mineração destacaram-se algumas redes envolvendo os seguintes nós: classes de preço de fábrica, laboratórios, e tarja. O relacionamento entre as variáveis tarja e laboratórios, Figura 3, gerou uma rede que sugere uma predileção dos laboratórios pela produção de medicamentos de tarja vermelha, que são vendidos apenas sob a prescrição de médicos ou dentistas. Entre os principais medicamentos de tarja vermelha comercializados no Brasil estão os fármacos para o tratamento de diabetes, hipertensão e medicamentos psicotrópicos.

Na mesma rede nota-se um alto número de registros onde a tarja do produto não é informada, além de observar 11 laboratórios que não informaram a tarja de nenhum dos seus medicamentos produzidos. Além disso, nenhum dos laboratórios produz apenas medicamentos de tarja preta, diferente do que ocorre com as tarjas vermelha e venda livre.

Quando relacionamos as variáveis faixas de preço de fábrica e laboratórios, Figura 4, observou-se que a maioria dos laboratórios tem em seu catálogo produtos cujo preço para os varejistas custa menos do que R$1000, sendo a faixa *Menos que R$100* a maior parcela entre todas. O número de laboratórios vai diminuindo conforme aumenta-se o valor dos produtos, chegando ao extremo onde apenas um laboratório produz medicamentos com valor de fábrica acima de 1 milhão de reais. Nesta mesma rede, também foi possível observar uma significativa quantidade de laboratórios que produzem apenas medicamentos com valor de fábrica abaixo de R$100. Ainda, dois dos 268 laboratórios produzem apenas medicamentos com custo acima de R$50000 e abaixo de 1 milhão de reais.

A rede monopartida, Figura 5, formada por nós da Classe Terapêutica por meio de projeção bipartida com as variáveis Faixa de Preço de Fábrica e Regime de Preço possui 534 nós, equivalentes às diferentes classes terapêuticas. Ao realizar o cálculo de modularidade, foram gerados 5 agrupamentos de classes terapêuticas identificados na rede por cores:

- Marrom: 224 nós
- Azul: 115 nós
- Verde: 97 nós
- Roxo: 40 nós
- Amarelo: 36 nós
- Vermelho: 22 nós

Destaca-se, por exemplo, que o grupo verde é composto por 60% de classes terapêuticas de venda livre (tarja), enquanto que o segundo grupo com maior proporção desta categoria de tarja conta com apenas 10%. Com relação ao tipo de produto, o agrupamento verde também é o que possui o maior número de fitoterápicos, com 8,32%, enquanto que em todos os outros grupos esta porcentagem não chega a 1%. Ainda, no grupo marrom há aproximadamente o dobro ou mais de classes terapêuticas com restrição hospitalar quando comparado aos outros agrupamentos.

O *dashboard* gerado com o software Power BI Desktop, Figura 6, permitiu observações amplas da base de dados como, por exemplo, o número total de laboratórios, produtos e substâncias e o cálculo, por exemplo, do valor médio do preço de fábrica de um subconjunto específico de dados. O painel possibilitou a segmentação precisa da base de dados, como a escolha de um ou vários laboratórios e o filtro de tarja e de regime de preço. Ainda, foi possível visualizar em gráfico de barras a comparação entre o preço de fábrica e o preço de revenda dos produtos, a distribuição de produtos por faixa de preço e a quantidade de produtos em cada categoria de tarja. No total, observou-se que a base de dados com 268 laboratórios tem registro de 6242 produtos e 2307 substâncias, sendo que a média do preço de fábrica entre todos

manipulação de dados com o propósito de gerar, por exemplo, painéis interativos. Disponível em https://powerbi.microsoft.com/.

os produtos é de aproximadamente R$3000. Nota-se também que a proporção de produtos entre as faixas de preço se manteve independentemente da tarja.

5 Conclusão ou Considerações Finais

Dentre as categorias de tarja, a tarja vermelha é a que é produzida pelo maior número de laboratórios, o que pode se relacionar com um cenário em que a hipertensão arterial é uma das principais causas de morte no Brasil, que acomete 24% das pessoas com mais de 18 anos, e em que a diabetes ocorre em 7% da população. Ambas as condições clínicas são tratadas por remédios controlados de tarja vermelha, em que é obrigatória a prescrição médica. Outras condições de saúde com diferentes prevalências também podem contribuir para o aumento da demanda de medicamentos de tarja vermelha.

Muitos registros de produtos não incluíram a informação de tarja, o que prejudica a análise assertiva da base de dados. Imagina-se que a não inclusão desse tipo de informação é devida à falhas humanas no processo de alimentação da base de dados. Essa situação ilustra um problema grave da gestão de dados e, se tratando de dados públicos, mantidos por uma instituição pública e relacionados com uma temática tão cara à população, calcula-se que o prejuízo gerado pela má gestão dos dados seja potencialmente alto.

Alguns laboratórios se reservam a produzir apenas um tipo específico de tarja, embora nenhum dos 268 laboratórios produza exclusivamente medicamentos de tarja preta. Os medicamentos de tarja preta são os psicoativos que têm um alto potencial de causar dependência, como as morfinas e anfetaminas e, por isso, deduz-se que seu uso restrito limite a demanda destes no mercado, talvez impossibilitando um modelo de negócios farmacêutico com produção exclusiva.

A maioria das classes terapêuticas registradas na base de dados tem o preço de fábrica girando em torno de R$100, enquanto que apenas duas classes terapêuticas tem custo acima de 1 milhão de reais. São elas *M5X - TODOS OS OUTROS FÁRMACOS COM AÇÃO MÚSCULO-ESQUELÉTICA* e *S1X1 - OUTROS PRODUTOS OFTALMOLÓGICOS SISTÊMICOS.* Ambas as classes terapêuticas são produzidas por um único laboratório, Novartis Biociências S.A., que produz todos os produtos com valor acima de um milhão de reais.

As relações intracluster da rede monopartida ainda precisam ser investigadas de maneira mais minuciosa a fim de entender de maneira plena quais as características fundamentais de cada um dos agrupamentos e a sua relevância para esta pesquisa

Portanto, o objetivo da pesquisa foi parcialmente alcançado uma vez que algumas relações entre as variáveis foram descobertas e investigadas. Contudo, é necessária a continuação do esforço de trabalho para analisar possíveis relações ainda não descobertas na base de dados manipulada neste estudo.

Referências

ALBERT, Réka; BARABÁSI, Albert-László. Statistical mechanics of complex networks. **Reviews of modern physics**, v. 74, n. 1, p. 47, 2002.

BARABÁSI, Albert-László. **Linked: The new science of networks**. 2003.

BÁRRIOS, Maria João; MARQUES, Rita; FERNANDES, Ana Alexandre. Aging with health: aging in place strategies of a Portuguese population aged 65 years or older. **Revista de Saúde Pública**, v. 54, 2020.

BATISTA, Gustavo Enrique de Almeida Prado et al. **Pré-processamento de dados em aprendizado de máquina supervisionado**. 2003. Tese de Doutorado. Universidade de São Paulo.

COSTA, Claudio Napolis et al. Descoberta de conhecimento em bases de dados. **Revista Eletrônica: Faculdade Santos Dumont**, v. 2, p. 20, 2019.

FAYYAD, Usama; PIATETSKY-SHAPIRO, Gregory; SMYTH, Padhraic. The KDD process for extracting useful knowledge from volumes of data. **Communications of the ACM**, v. 39, n. 11, p. 27-34, 1996.

HAND, David J. Principles of data mining. **Drug safety**, v. 30, n. 7, p. 621-622, 2007.

METZ, Jean et al. Redes complexas: conceitos e aplicações. 2007.

WASSERMAN, Stanley et al. Social network analysis: Methods and applications. 1994.

Apêndice A

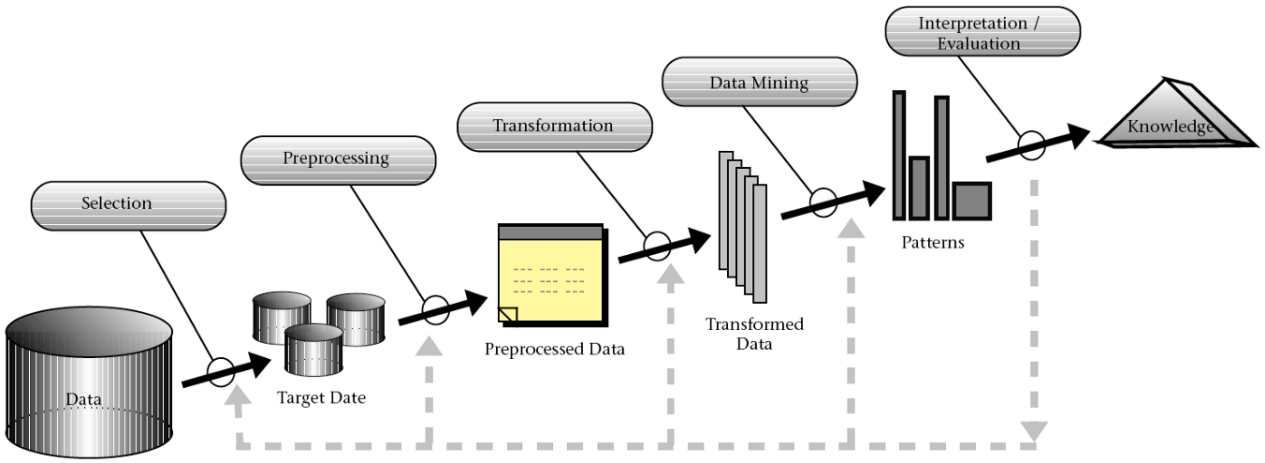

Figura 1 - Etapas do processo KDD.
Fonte: Fayyad et al. (1996).

```
graph [
directed 0


node [ id {{jsonize(cells.classeTerapeutica.value)}}
variavel "Classe Terapeutica" agrupamento "grupo A" ]

node [ id {{jsonize(cells.faixaPrecoFabrica17.value)}}
variavel "Faixa de Preço de Fabrica" agrupamento "grupo
B" ]

node [ id {{jsonize(cells.regimeDePreco.value)}}
variavel "Regime de Preço" agrupamento "grupo B" ]


edge [ source {{jsonize(cells.classeTerapeutica.value)}}
target {{jsonize(cells.faixaPrecoFabrica17.value)}}]

edge [ source {{jsonize(cells.classeTerapeutica.value)}}
target {{jsonize(cells.regimeDePreco.value)}}]

]
```

Figura 2 - Código de mapeamento GML pelo software OpenRefine para geração de rede tripartida.
Fonte: autoria própria.

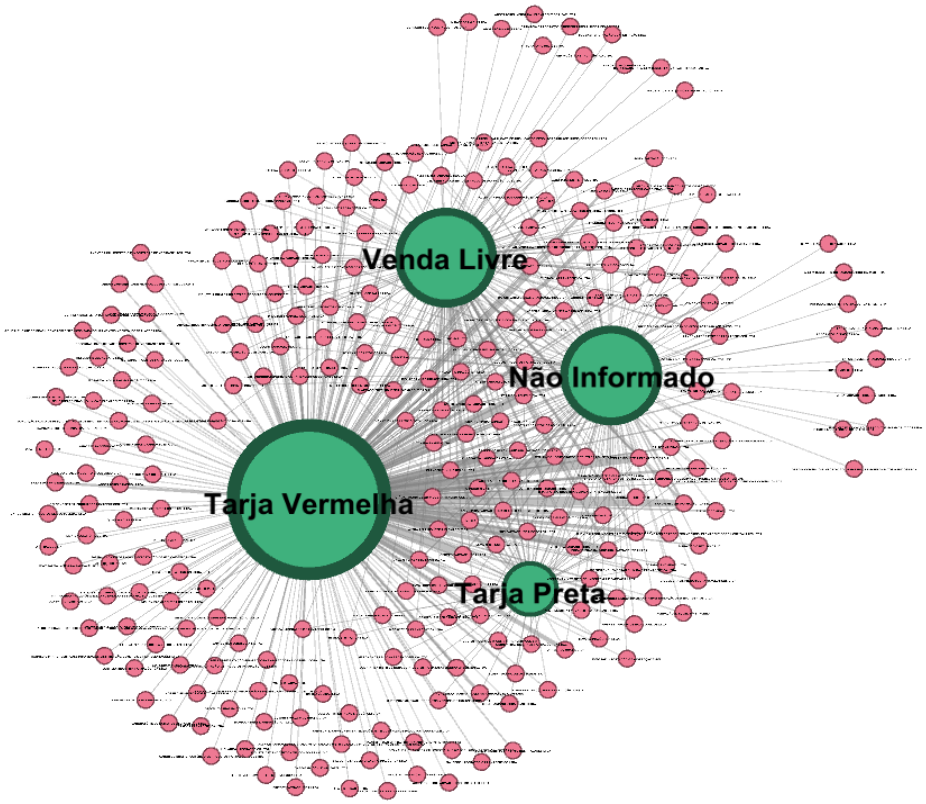

Figura 3 - Rede informacional bipartida relacionando as variáveis tarja e laboratório.
Fonte: autoria própria, com apoio do software Gephi.

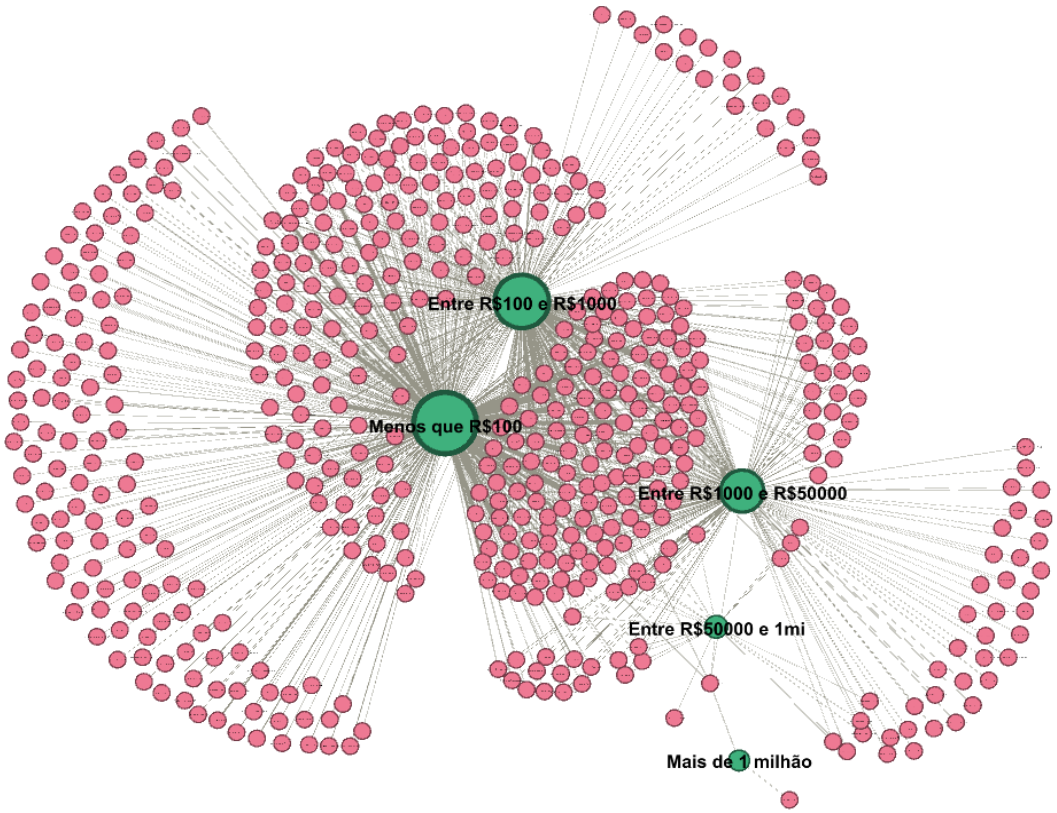

Figura 4 - Rede informacional bipartida relacionando as variáveis faixa de preço de fábrica e classe terapêutica.
Fonte: autoria própria, com apoio do software Gephi.

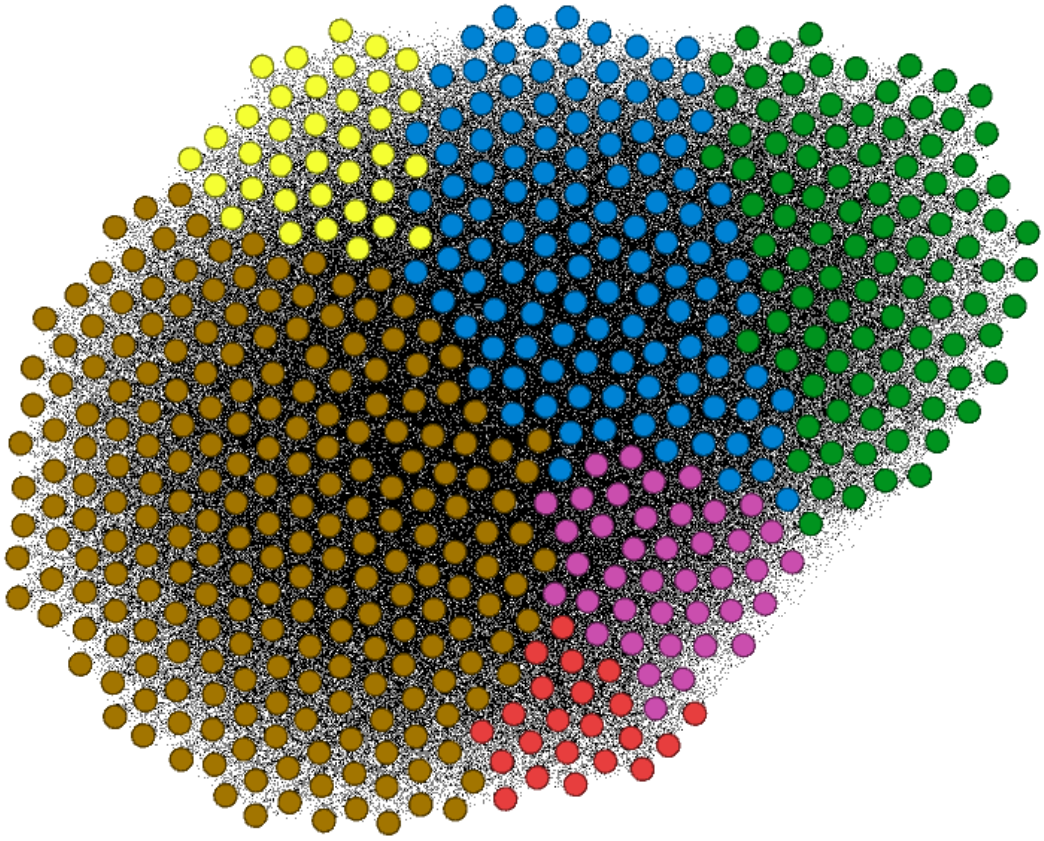

Figura 5 - Rede informacional monopartida relacionando as variáveis faixa de preço de fábrica, classe terapêutica e regime de preço colorida após cálculo de modularidade.
Fonte: autoria própria, com apoio do software Gephi.

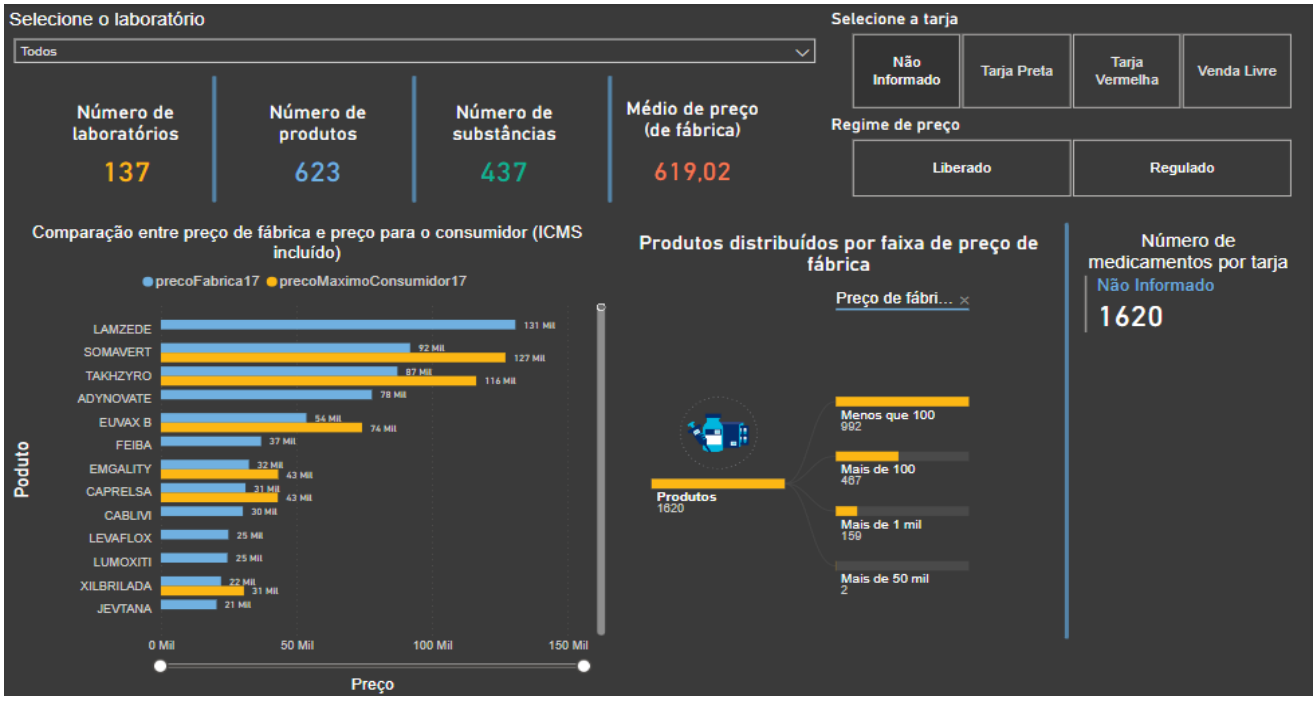

Figura 6 - Dashboard de variáveis da base de dados
Fonte: autoria própria, com apoio do software Power BI Desktop