Exploração Semântica de Comorbidades da COVID-19: Ferramentas para predição de mortes.

Felipe Calmon, Fernanda Baião e Lucas Maddalena

Departamento de engenharia industrial, Pontifícia Universidade Católica do Rio de Janeiro, Rua Marquês de São Vicente 225, Rio de Janeiro, 22451-900, Brasil.

Abstract

A ontologia na ciência de dados é uma ferramenta que auxilia formalmente a estrutura de um sistema, na identificação das suas entidades e as suas relações entre si. Dessa forma, em um mundo globalizado, onde os dados são produzidos massivamente, entretanto, em muitas esferas sem padronizações ou em organizações que não necessariamente exprimem a real semântica da realidade, a ontologia, se apresenta como excelente alternativa para o encadeamento e gerenciamento logico dos dados. Posto isto, o seguinte trabalho teve por objetivo avaliar um conjunto de dados provenientes do SIVEP-Gripe, sob o enriquecimento semântico da Ontologia de Doenças Humanas (DO), e examinar as benesses desse artifício as mais variadas técnicas de Machine learning empregadas no processo de pesquisa. Em resumo, todo esse processo objetiva na predição de morte dos pacientes provenientes do SIVEP-Gripe e busca compreender como esse conjunto de doenças ou não, se correlatam com esses óbitos.

Keywords

Ontologia de doenças, Clustering, COVID-19

1. Introdução

O historial da COVID-19 se principia na sua descoberta em dezembro de 2019 até a sua seriação como pandemia pela Organização Mundial da Saúde (OMS) em março de 2020. No primeiro momento, identificado apenas na cidade de chinesa de Wuhan, o novo coronavírus, chamado SARS-CoV-2, rapidamente tomou proporções globais, revertendo-se em uma emergência de saúde pública internacional declarada pela OMS em janeiro de 2020, e em seguida declarando-se como uma pandemia em março do mesmo ano. Posto isto, após a afirmativa da doença, foi imperativo a coordenação nacional para combater os avanços da doença, e mitigar os riscos do vírus sobre os mais diversos tecidos sociais do globo [Fonte].

As Ontologias são eximiras aliadas da ciência de dados, visto que, proporcionam uma estrutura semântica que consente uma compreensão mais profunda e precisa dos dados ao definir relações e classificações para um domínio específico. Dessa forma, elas ofertam uma base comum de conhecimento compartilhado, e por consequência, facilitam a integração de

* Corresponding author.
† These authors contributed equally.
aleksandr.ometov@tuni.fi (A. Ometov); t.princesales@utwente.nl (T. P. Sales); manfred.jeusfeld@acm.org (M. Jeusfeld)
0000-0003-3412-1639 (A. Ometov); 0000-0002-5385-5761 (T. P. Sales); 0000-0002-9421-8566 (M. Jeusfeld)

dados de diversas fontes e a interoperabilidade entre sistemas. Ao que tange, a COVID-19, diversos trabalhos já propuseram a aplicação de ontologias, como Fonte [Fonte].

O trabalho a seguir, manipula os dados provenientes do SIVEP-Gripe sobre as comorbidades de pacientes diagnosticados com a COVID-19 no Estado do Rio de Janeiro. O propósito substancial é observar a implicação que a abordagem semântica permiti ao conseguir predizer ou não o óbito dos pacientes e por conseguinte averiguar o subconjunto de doenças que mais se assemelham a esse obituário. Para isso, após a aplicação de técnicas de clustering foi preciso separar o conjunto de dados em dois casos, o primeiro caso, sem o enriquecimento semântico, onde cada hospitalização representava um vetor binário, e após a submissão da métrica de similaridade entre os cossenos (A definir).

Ademais, no que tange ao segundo caso, o domínio de ontologia escolhido foi a DO (Ontologia de Doenças), que é um recurso elementar na integração entre vocabulários médicos, facilitando a associação das comorbidades em análise com os conceitos da DO. Para o cálculo da similaridade, foi empregado a métrica que usa os termos da DO, onde cada paciente pode apresentar mais de uma comorbidade, logo, é preciso que a avaliação da métrica seja conjunta, para que a comparação seja entre subconjuntos e não sobre termologias individuais. Deste modo, por permitir uma análise mais robusta, se consideramos hipoteticamente um conjunto de doenças como “Hipertensão essencial, Diabetes mellitus, Doença cerebrovascular, Doença renal crônica”, a métrica empregada embasa-se no cálculo de similaridades em grupo entre termos na DO, promovendo uma compreensão mais profunda e contextualizada das relações entre as doenças.

Em resumo, a abordagem semântica de forma consciente quando de encontro a subconjuntos semelhantes de hostilizações, é avaliada na etapa de Pós-Processamento de Dados usando métricas de qualidade de cluster. Prontamente, será realizada a previsão de riscos dos usuários e a observação das doenças mais prevalentes nesse processo. A investigação na parte final se inclina na busca de demonstrar a eficácia dos clusters resultantes de cada cenário na segmentação dos subconjuntos de doenças e na sua capacidade encontrar mecanismos de predição óbitos de seus usuários, pois a importância da abordagem semântica na identificação de padrões e tendências é relevante para a saúde pública.

2. Ontologia e DO (Diseiase Ontology)

A ontologia é interpelada de duas maneiras distintas: A primeira abordagem, depreende do ponto de vista filosófico, envolve o estudo da natureza e estrutura da realidade; já o segundo aspecto, na ciência da computação, trata-se de um instrumento que modela a estrutura de um sistema [fonte]. Nesse sentido, a Disease Ontology (DO) é um eixo de conhecimento que centraliza informações sobre doenças humanas, fornecendo um guia para análises biomédicas. Portanto, o seu designo é simplificar a integração de dados genômicos e clínicos, ofertando uma estrutura precisa e compreensível para a investigação e entendimento das doenças. A título de exemplo, no diagrama abaixo Figura 1, a ‘doença do coração’ é uma doença do ‘sistema cardiovascular’, entretanto {cardiomiopatia, doença

congênita do coração, aneurisma do coração e câncer do coração} é uma ‘doença do coração’. A Disease Ontology, sob a liderança de Lynn Schriml, da Escola de Medicina da Universidade de Maryland, visa unificar dados diversos e promover uma visão completa das enfermidades humanas [Fonte].

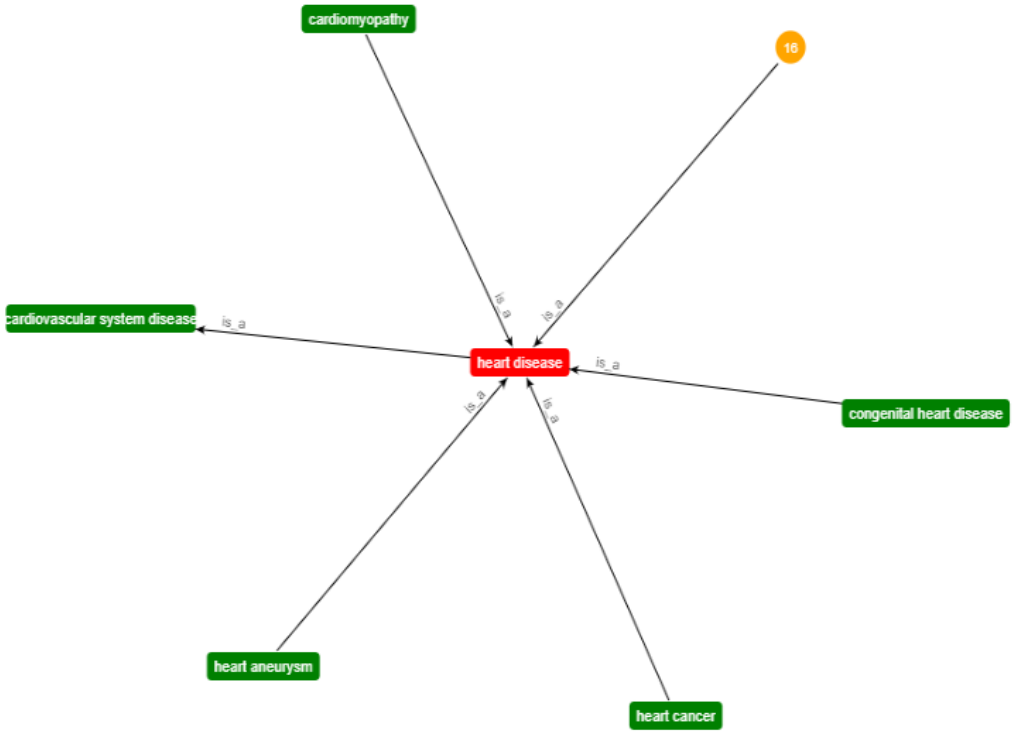

Figura 1: Diagrama de relações Ontológicas. Fonte: (https://disease-ontology.org/do).

3. Ciclo de vida na ciência de dados e os benefícios da semântica ontológica

O ciclo de vida na ciência de dados pode ser entendido como o processo que guia projetos desde a compreensão do problema até o feedback das ações implementadas. Posto isto, é plausível destacar algumas fases pelos quais se compreendem nesse ciclo [fonte].

- **Coleta de dados:** Consiste em obter os dados corretos, considerando a complexidade do mundo real e as diferentes perspectivas dos indivíduos.
- **Pré-processamento de Dados:** Nesta etapa, os dados são analisados e manipulados para serem tratados e transformados adequadamente, preparando-os para a etapa seguinte de Mineração de Dados. Isso envolve lidar com valores ausentes, atípicos e selecionar e organizar características relevantes.
- **Mineração de Dados:** Nessa fase, as técnicas computacionais e estatísticas são aplicadas para extrair conhecimento dos dados. Ou seja, isso inclui Aprendizado de Máquina supervisionado e não supervisionado, que se vale de algoritmos de

classificação, regressão e agrupamento para descobrir padrões nos dados e prever as características desses tecidos em análise.

- **Pós-processamento de dados:** Após a etapa da Mineração de Dados, os resultados são interpretados, reflexionando a sua qualidade e considerando possíveis vieses. Os resultados são utilizados para reavaliar o modelo em uso e a avaliação pode ser feita usando métricas como o Coeficiente de Silhueta, por exemplo.

O uso de ontologias é crucial na ciência de dados, oferecendo uma estrutura semântica que aprofunda a compreensão dos dados ao definir relações e classificações para um domínio específico. Na análise da COVID-19, ontologias têm sido aplicadas em diversos estudos. Por exemplo, um trabalho que utiliza dados do SIVEP-Gripe sobre comorbidades de pacientes com COVID-19 no Rio de Janeiro buscando predizer óbitos e identificar as doenças mais associadas a esse cenário. À vista disso, com o emprego das técnicas de clustering, os dados são comparados em dois cenários: O primeiro, sem enriquecimento semântico e outro com a aplicação da Ontologia de Doenças (DO). Desta maneira, a última abordagem permite uma análise mais robusta, considerando um agrupamento de doenças, o que aprimora a compreensão das relações entre elas. Por fim, na etapa de Pós-Processamento de Dados, métricas de qualidade de cluster são operadas para avaliar a eficácia da abordagem semântica na segmentação de doenças e na previsão de óbitos, todo o processo descrito anteriormente pode ser enquadrado na Figura 2. Essa análise é fundamental para a saúde pública, destacando a importância da abordagem semântica na identificação de padrões e tendências [Fontes].

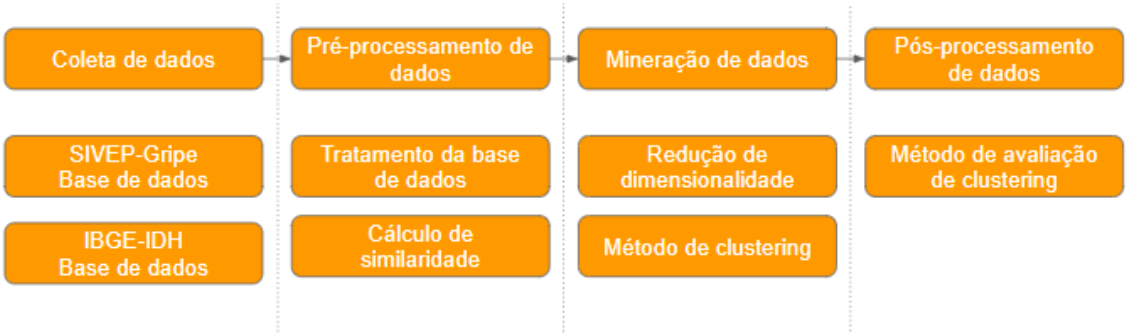

Figura 2: O Ciclo da ciência de dados aplicado ao estudo em questão. Fonte: Autor

4. Coleta e análise de dados

O SIVEP-Gripe foi criado no ano 2000 para monitorar o vírus da gripe no Brasil, expandindo-se em 2009 para monitorar a SRAG (Síndrome Respiratória Aguda Grave). Por conseguinte, com o advento da pandemia da COVID-19 em 2020, houve uma adaptação do sistema para incluir o novo coronavírus, projetando fortalecer a vigilância de vírus respiratórios. Esta adaptação foi feita pela Secretaria de Vigilância em Saúde do Ministério da Saúde, direcionando o Sistema Nacional de Vigilância em Saúde para lidar com a circulação simultânea do SARS-CoV-2, Influenza e outros vírus respiratórios durante a emergência de saúde pública [Fonte]. Ademais, os dados do Índice de Desenvolvimento Humano (IDH) dos municípios brasileiros foram angariados do Instituto Brasileiro de

Geografia e Estatística (IBGE), uma procedência de dados confiável e reconhecida para informações estatísticas e indicadores socioeconômicos referentes ao tecido social brasileiro [Fonte].

5. Pré-Processamento dos dados

Na seção de Pré-processamento de dados, o cerne do conteúdo estará na preparação inicial dos dados para análise posterior, mineração de dados. Dessarte, esta etapa é crucial para garantir a qualidade e a relevância dos dados, pois dentro dessa seção, as atividades serão bifurcadas em duas subseções principais: o tratamento de dados, que envolve a limpeza, organização e transformação dos dados brutos em um formato adequado para análise, e o enriquecimento semântico, que se concentra em adicionar informações provenientes da Disease Ontology (DO), para potencializar a sua utilidade e interpretação [Fonte].

5.1 Tratamento da base de dados

O processamento dos dados teve por início a importação das bibliotecas essenciais, tais como pandas, numpy e matplotlib. Em seguida, o arquivo CSV contendo os dados dos pacientes foi então lido e atribuído ao DataFrame “srag_bruto”, com categorias de dados singulares definidos para algumas colunas. Nessa lógica, a coluna de datas foi convertida para o tipo datetime e o DataFrame foi filtrado para incluir apenas as datas dentro de um intervalo característico de tempo. Os pacientes com evolução "Ongoing", em andamento, foram removidos e uma nova coluna 'IS_DEATH' foi criada para indicar a ocorrência de óbito ou não do enfermo.

O Data Frame final dispôs de colunas selecionadas criteriosamente, o que incluiu bases como sexo, idade, comorbidades e localização geográfica. Posto isto, colunas que continham valores "Yes" foram convertidas para 1 e os demais valores para 0, para simbolizar os dados binários. Após esta etapa, o DataFrame estava preparado para a próxima fase, que consistiu na conversão de valores categóricos em binários, como na coluna "CS_SEXO", Tabela 1.

Tabela 1
Caracteristica dos dados. Fonte: Autor

Classificação	Aparência	Comentarios
Dados binários	0 ou 1	O dado binário é um condicionamento para a representação dos dados em que as informações podem assumir apenas dois valores distintos, geralmente representados como 0 e 1.
Dados categóricos	Diferentes atributos dado a um objeto (Azul,	O dado categórico é uma forma de dados em que as informações são divididas em categorias ou grupos distintos e não ordenados.

verde, amarelo e
etc)

Adicionalmente, foram geradas colunas adicionais para significar os diferentes grupos étnicos com alicerce na coluna "CS_RACA". Além do mais, outras etapas envolveram o escalonamento dos valores da coluna "NU_IDADE_N" usando MinMaxScaler e a importação de um novo DataFrame contendo o Índice de Desenvolvimento Humano (IDH) dos municípios brasileiros objetivando a criação da coluna “IDH_MUN_RES” e IDH_MUN_INTE, Figura 3.

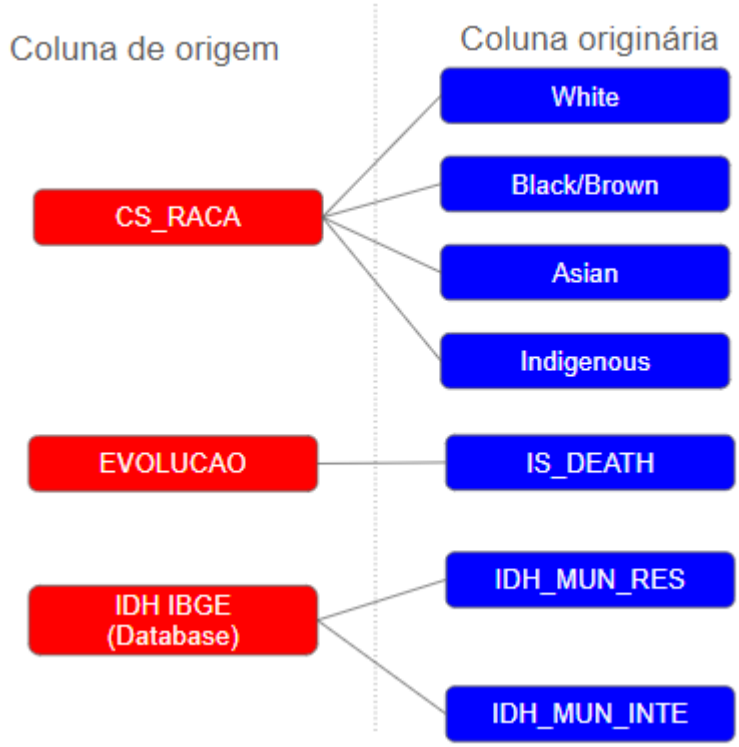

Figure 3: Origem das novas colunas. Fonte: Autor

5.2 Enriquecimento semântico

Após a etapa de limpeza dos dados, da base de dados bruta derivada do SIVEP-Gripe, foi efetuado o processo de identificação das colunas relevantes ao enriquecimento semântico, no estudo em questão, as colunas que implicam nas comorbidades dos pacientes. Prontamente, um dicionário foi empregado para renomear as colunas selecionadas, atribuindo a cada uma um identificador único correspondente ao vocabulário da Ontologia de Doenças (DOID), Tabela 2. Esse estágio foi basilar para estabelecer uma estrutura semântica que possibilitasse uma compreensão mais precisa dos dados.

Tabela 2

Correspondência das comorbidades na disease ontology. Fonte: [Fonte]

Código DO	Comorbidade	Correspondente DO
DOID:114	Cardiopatia	Heart Disease
DOID:2531	Doença Hematológica	Hematologic cancer
DOID:14250	Síndrome de Down	Down syndrome
DOID:409	Doença Hepática	Liver disease
DOID:2841	Asma	Asthma
DOID:0110741	Diabetes melitus	Type 1 diabetes mellitus 2
DOID:3230	Doença Neurológica	High pressure neurological syndrome
DOID:850	Pneumopatia crônica	Lung Diseiase
DOID:471	Imunodeficiencia	Autoimmune disease
DOID:557	Doença Renal crônica	Kidney disease
DOID:9970	Obesidade	Obesity

Após a renomeação das colunas não fundamentais para aquele período em análise, ou seja, apenas aquelas contendo informações sobre a presença ou ausência de determinadas doenças no paciente, foi desenvolvida uma função para renomear as linhas do DataFrame, substituindo os índices numéricos por uma lista ordenada das doenças presentes em cada registro.

Com a formação do novo arquivo CSV contendo as comorbidades indexadas aos códigos da Ontologia de doenças, os dados foram usados na ferramenta Rstudio com a prerrogativa de substituir os valores das linhas numéricas no DataFrame principal pelos valores correspondentes de similaridade. Dessa forma, o processo foi crucial para enriquecer os dados com informações semânticas entre os conjuntos pré-determinados, permitindo uma análise mais aprofundada e contextualizada das relações entre as doenças.

Tabela 2

Correspondência das comorbidades na disease ontology. Fonte: [Fonte]

	DOID:114	DOID:2531
DOID:14250,DOID:2841,DOID:850	0.109892	1.000000
DOID:114,DOID:2531,DOID:2841	0.226732	0.071430

Em síntese, o DataFrame decorrente do enriquecimento semântico foi reintegrado ao conjunto selecionado de colunas do DataFrame original, incluindo informações como sexo, idade e localização geográfica, entre outros fatores relevantes. Portanto, essa etapa consumiu o processo de tratamento e enriquecimento dos dados, preparando-os para a mineração dos dados.

6. Mineração dos dados

No que tange a etapa da mineração dos dados, o foco estará na aplicação de técnicas computacionais e estatísticas, tencionando extrair conhecimento dos dados. Dentro dessa seção, as atividades serão orquestradas em duas dimensões fundamentais: a redução de dimensionalidade, que engloba a seleção das variáveis mais relevantes e a redução da complexidade dos dados em questão, e o método de clustering, que tem por designo agrupar os dados em conjuntos semelhantes, proporcionando a identificação de padrões e tendências.

6.1 Redução de dimensionalidade

A gênese da mineração dos dados teve início com a normalização, uma ação prática que tem por intenção a distribuição mais homogênea e facilitação da análise. Por conseguinte, foi preciso a aplicação da análise dos componentes principais, que consiste em uma proposta que tem por finalidade básica, a análise dos dados analisados pretendendo a sua redução, através da eliminação de sobreposições e a escolha dos modelos mais representativos de dados por intermédio de combinações lineares das variáveis originais, Figura 4 [Fonte].

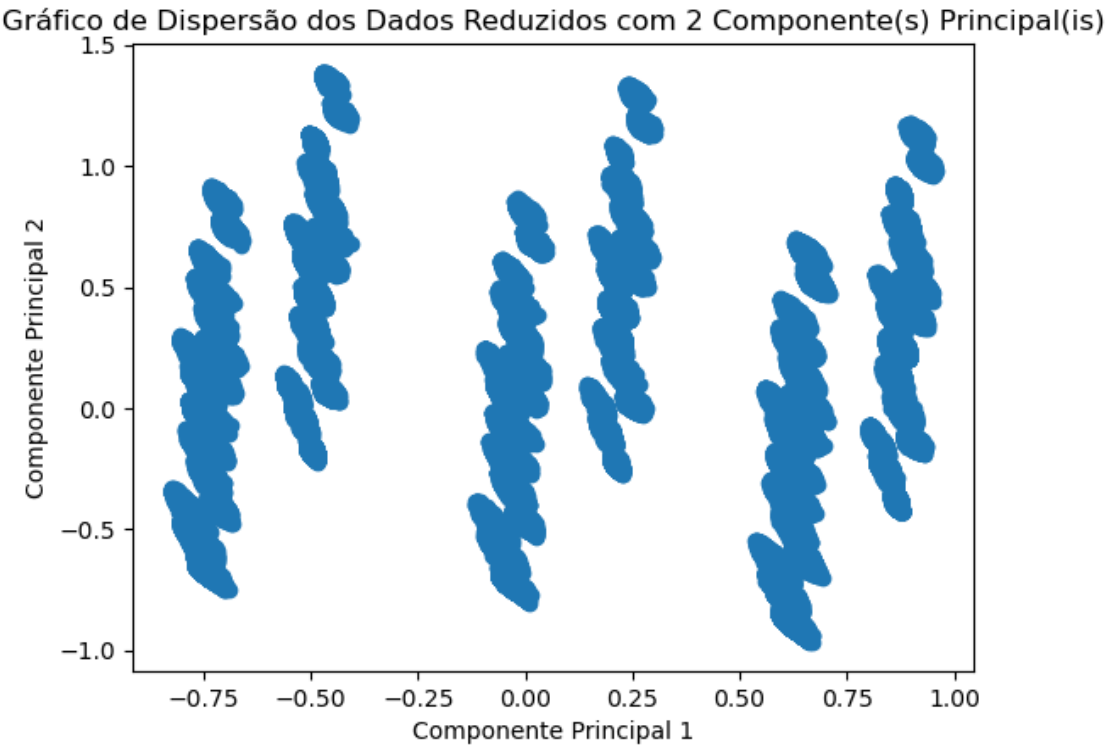

Figura 4: Expressão gráfica do teste de PCA para duas dimensões. Fonte: Autor

Desse modo, para uma compreensão mais elucidativa de como cada componente contribui para a variabilidades dos dados, os dados resultantes do PCA são submetidos ao método do cotovelo, avaliando a variância explicada cumulativa em relação ao número de componentes. Posteriormente, foram produzidos gráficos de dispersão dos dados reduzidos para duas dimensões, o que permitiu a visualização da distribuição dos registros em relação aos componentes principais. Por fim, essa visualização facilitou a identificação de padrões e agrupamentos nos dados, Figura 5.

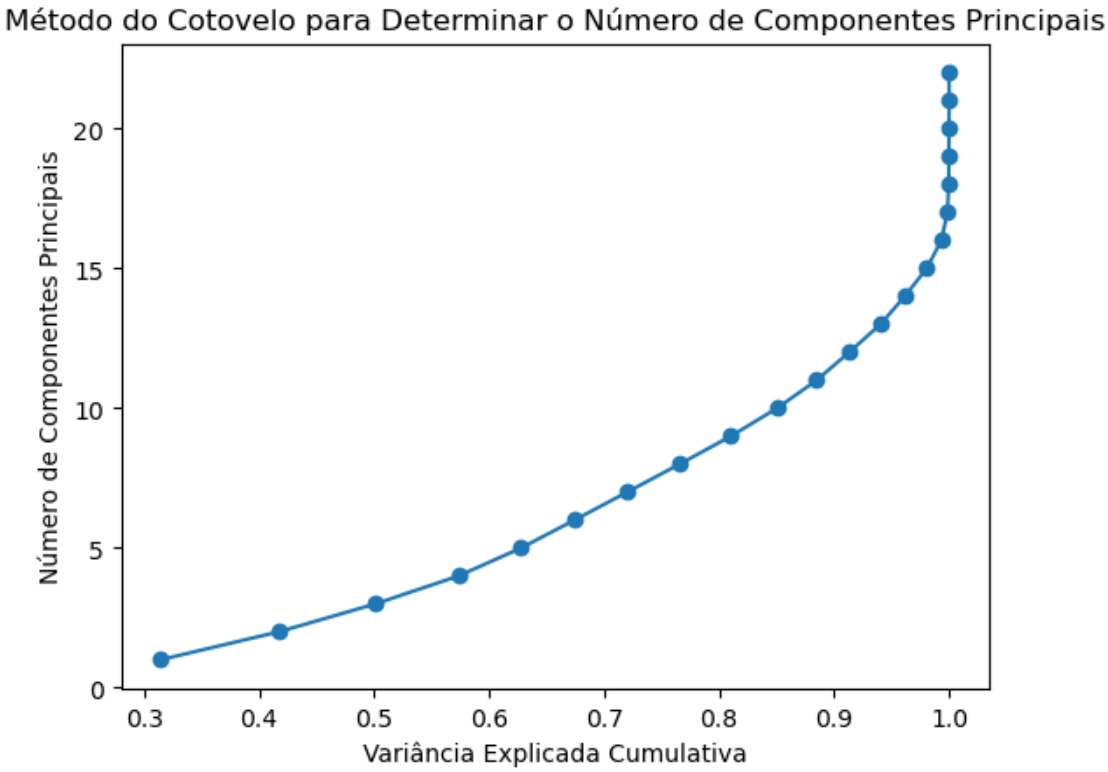

Figura 5: Expressão gráfica do método de cotovelo. Fonte: Autor

6.2 Método de clustering

Na pretensão de formalizar agrupamentos conforme a similaridade dos dados reduzidos, foram testados diferentes métodos, como o KMeans e o GMM, Tabela 4, e em seguida foram gerados gráficos e um Data Frame para visualizar os agrupamentos formados pelos métodos, oferecendo insights visuais sobre a estrutura dos dados, Figuras 6 e 7.

Tabela 4

Definição dos métodos de clustering escolhidos. Fonte: [Fonte]

Método	Descrição
K-means (KMeans)	É um algoritmo, que funciona atribuindo pontos de dados a um dos K clusters com base na proximidade

	ao centroide mais próximo. O algoritmo itera entre atribuir pontos aos clusters e recalculando os centroides até que a variação intra-cluster seja minimizada.
Gaussian Mixture Model - GMM	É um método que assume que todos os pontos de dados são gerados a partir de uma mistura de várias distribuições gaussianas. Ele estima as distribuições gaussianas subjacentes aos dados e suas probabilidades de ocorrência em cada ponto de dados. Isso permite modelar a estrutura complexa dos dados, incluindo clusters de diferentes formas e tamanhos.

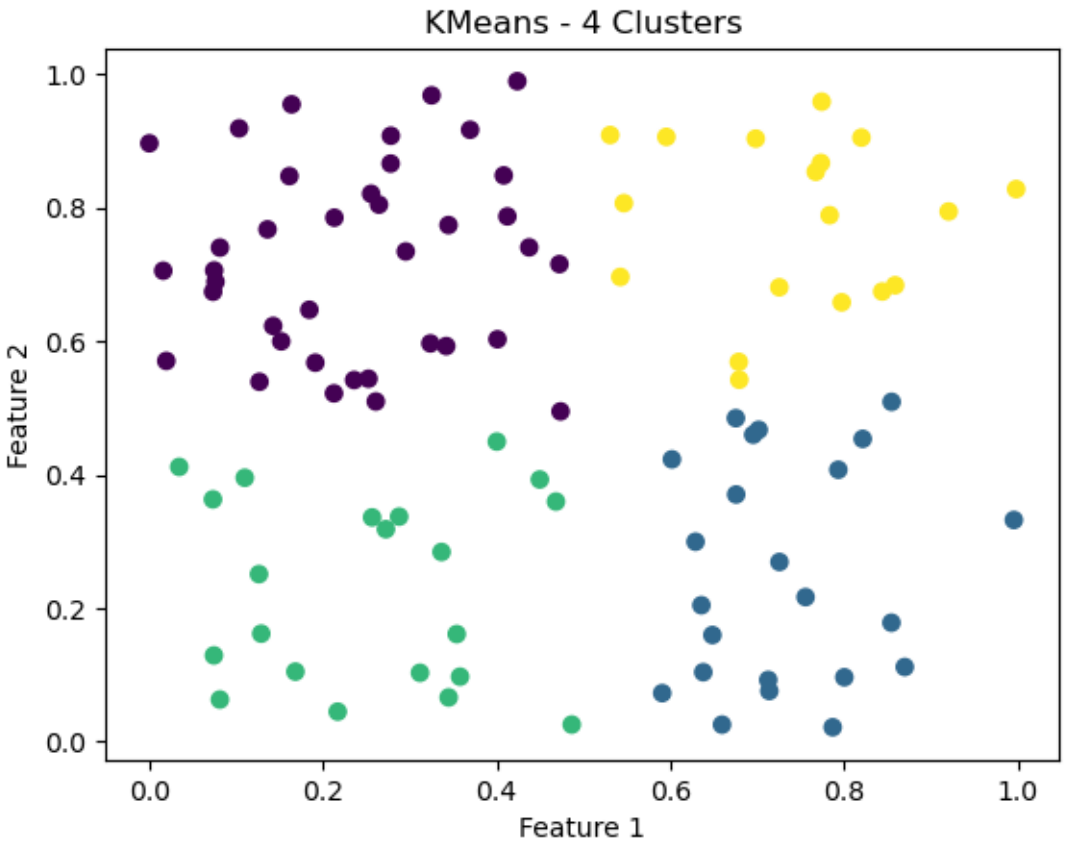

Figura 6: Uso do método de clustering Kmeans para 4 clusters. Fonte: Autor

	Number of Clusters	KMeans (BIC)	KMeans (DBI)	GMM (BIC)	GMM (DBI)
0	4	228002.311794	0.767031	3.041364e+06	0.939061
1	5	175349.931713	0.764013	2.689160e+06	0.856164
2	6	132422.853596	0.679400	-9.930712e+04	1.410915
3	7	103449.654243	0.673697	-1.357853e+05	1.313389
4	8	77681.495392	0.641218	-2.744522e+05	1.210805
5	9	60450.990987	0.574258	5.445308e+04	1.134005
6	10	52944.890128	0.597649	-4.785445e+05	1.309552
7	11	45418.861502	0.631389	-5.777862e+05	1.070026
8	12	40460.688930	0.617441	-6.698667e+05	1.003908
9	13	35805.098967	0.636928	-1.822224e+06	0.927771
10	14	32197.292241	0.628324	-1.803802e+06	0.960324
11	15	29472.364737	0.671687	-1.865058e+06	0.860722

Figura 7: Data Frame com a avaliação dos clusterings formados por cada método. . Fonte: Autor

7. Pós-processamento de dados

Na etapa de pós-processamento de dados no ciclo de ciência de dados, o foco se volta para a revisão e refinamento dos resultados obtidos após análise e modelagem, na mineração de dados. Nesta fase basilar, os dados analisados foram interpretados, visando possíveis ajustes e percepções foram consolidadas para um veredito final, nessa circunstância, a predição de morte dos pacientes com COVID. Ou seja, este processo de validação e comunicação é primordial para garantir a precisão e a relevância dos resultados obtidos durante toda a análise de dados anterior.

7.1 O Uso da ferramenta XGBoost

A ferramenta XGBoost pode ser compreendida como uma biblioteca de código aberta amplamente utilizada para aprendizado de máquina fundamentado em árvores de decisão. Portanto, a sua importância reside na potencialidade de oferecer alta eficiência computacional e precisão preditiva, nesse evento a predição de mortes por COVID, em uma gama de problemas de regressão e classificação [Fonte].

Para o início do delineamento dessa etapa, o código foi bifurcado em dois conjuntos de modelos XGBoost para prever mortes em pacientes, diferenciados pela seleção de características relevantes. Dessa forma, no primeiro conjunto, as características selecionadas são mais genéricas, com condicionamento médico dos pacientes sem enriquecimento semântico, enquanto no segundo conjunto, as características os dados das

comorbidades enriquecidas semanticamente. Nesse sentido, ambos os conjuntos seguiram a mesma metodologia: Os dados foram divididos em treinamento e teste, e assim um modelo XGBoost for treinado e por consequência, previsões seguidas de uma avaliação do desempenho do modelo foram geradas. Por fim, a saída é um relatório de classificação que mostra métricas como precisão, recall e F1-score para cada classe de previsão, Figura 8,9.

	precision	recall	f1-score	support
0	0.87	0.90	0.88	222089
1	0.79	0.74	0.76	112307
accuracy			0.84	334396
macro avg	0.83	0.82	0.82	334396
weighted avg	0.84	0.84	0.84	334396

Figura 8: Relatório de performance do XGBoost sem semântica.

	precision	recall	f1-score	support
0	0.86	0.89	0.88	222089
1	0.77	0.72	0.75	112307
accuracy			0.84	334396
macro avg	0.82	0.81	0.81	334396
weighted avg	0.83	0.84	0.83	334396

Figura 9: Relatório de performance do XGBoost com semântica.

7.2 Definição de importância das features

Após a confecção dos dados oriundos do modelo de predição, foi calculado e exibido a importância das características (features) nos exemplares. Sendo assim, os protótipos foram diferenciados pela seleção de características sem e com semântica. Desse modo, o código extrai a 'importância' das características dos modelos `model_01` e `model_02`, e em seguida, ele cria um dicionário que mapeia as características aos seus respectivos valores de importância. Em sequência, as características são ordenadas de acordo com sua importância e exibidas em ordem decrescente. Isso permite uma compreensão visual das características que mais influenciam nas previsões de morte em cada modelo.

References

[Incompleto e sem ajustes]

1. Guarino, N., Oberle, D., & Staab, S. (2009). What Is an Ontology? In S. Staab & R. Studer (Eds.), Handbook on Ontologies (pp. xx-xx). Springer-Verlag Berlin Heidelberg. DOI: 10.1007/978-3-540-92673-3
2. Amaral, G., Baião, F., & Guizzardi, G. Foundational Ontologies, Ontology-Driven Conceptual Modeling and their Multiple Benefits to Data Mining. (Autor).
3. Brebels, A., Shcherbakov, M. (2018). Lean Data Science Research Life Cycle: A Concept for Data Analysis Software Development. Comunicações em Ciência da Computação e Informação. DOI: 10.1007/978-3-319-11854-3_61.
4. Ministério da Saúde. (2021). Guia Rápido SIVEP-GRIPE [PDF]. Recuperado de https://www.saude.ba.gov.br/wp-content/uploads/2021/05/GUIA-RAPIDO-SIVEP-GRIPE-atualizado-em-maio_2021.pdf
5. Oliveira, L. A. P. de, & Simões, C. C. da S. (2005). O IBGE e as pesquisas populacionais. *Revista Brasileira de Estudos de População*, 22(2), 291-302.
6. Conci, A., & Oliveira, L. S. (s.d.). Principal Component Analysis: A Tutorial. Recuperado de http://www2.ic.uff.br/~aconci/PCA-ACP.pdf
7. scikit-learn. (s.d.). Gaussian Mixture Model (GMM). Recuperado de https://scikit-learn.org/stable/modules/mixture.html#gmm
8. scikit-learn. (s.d.). K-means. Recuperado de https://scikit-learn.org/stable/modules/clustering.html#k-means