Instituto Serzedello Corrêa
Escola Superior do Tribunal de Contas da União

**MARCELO CHAVES CURCIO**

# ANÁLISE DE CLUSTERIZAÇÃO NOS ESTABELECIMENTOS DE SAÚDE DO CNES

**Brasília**

**2020**

**MARCELO CHAVES CURCIO**

# ANÁLISE DE CLUSTERIZAÇÃO NOS ESTABELECIMENTOS DE SAÚDE DO CNES

Trabalho de conclusão do curso de pós-graduação *lato sensu* em Análise de Dados para o Controle realizado pela Escola Superior do Tribunal de Contas da União como requisito para a obtenção do título de especialista em Análise de Dados.

Orientador: Prof. MSc. Saul Campos Berardo

**Brasília**

**2020**

# REFERÊNCIA BIBLIOGRÁFICA

CURCIO, Marcelo Chaves. **Análise de Clusterização nos Estabelecimentos de Saúde do CNES.** 2020. Trabalho de Conclusão de Curso (Especialização em Análise de Dados) – Escola Superior do Tribunal de Contas da União, Instituto Serzedello Corrêa, Brasília DF.

# CESSÃO DE DIREITOS

NOME DO AUTOR: Marcelo Chaves Curcio
TÍTULO: Análise de Clusterização nos Estabelecimentos de Saúde do CNES
GRAU/ANO: Especialista/2020



_____

Marcelo Chaves Curcio
chavesc@tcu.gov.br

Ficha catalográfica

Curcio, Marcelo Chaves

Análise de Clusterização nos Estabelecimentos de Saúde do CNES / Marcelo Chaves Curcio; orientador, Saul Campos Berardo, 2020.

77 p.

Monografia (especialização) - Escola Superior do Tribunal de Contas da União, Curso de Especialização em Análise de Dados para o Controle, Brasília, 2020.

Orientações em:

1. Análise de Dados. 2. Mineração de Dados. 3. Clusterização. I. Berardo, Saul Campos. II. Escola Superior do Tribunal de Contas da União. Especialização em Análise de Dados para o Controle. III. Título.

**MARCELO CHAVES CURCIO**

# ANÁLISE DE CLUSTERIZAÇÃO NOS ESTABELECIMENTOS DE SAÚDE DO CNES

Trabalho de conclusão do curso de pós-graduação lato sensu em Análise de Dados para o Controle realizado pela Escola Superior do Tribunal de Contas da União como requisito para a obtenção do título de especialista.

Brasília, 9 de abril de 2020.

**Banca Examinadora:**

_____

Prof. MSc. Saul Campos Berardo

Orientador

Instituto Serzedello Corrêa - TCU

_____

Prof. Dr. Edans Flávius de Oliveira Sandes.

Instituto Serzedello Corrêa - TCU

*“Amamos mais o que conquistamos com sofrimento.”*
*Aristóteles (384 a.C. – 322 a.C.)*

*“O sucesso é ir de fracasso em fracasso sem perder entusiasmo.”*
*Winston Churchill (1874 – 1965)*

*“O importante não é vencer todos os dias, mas lutar sempre.”*
*Waldemar Valle Martins (1926 – 2004)*

# AGRADECIMENTOS

Agradeço a Deus em primeiro lugar, porque sem ele nada seria possível.

À minha esposa Nathalia e aos meus filhos Davi e Sarah pelo incentivo, compreensão e paciência demonstrados durante o período do projeto.

Aos meus pais Geraldo e Sylvia que sempre estiveram ao meu lado me apoiando ao longo de toda a minha caminhada.

Ao meu orientador Saul Campos Berardo pelo suporte, pela dedicação do seu escasso tempo ao projeto, pelas correções e incentivos.

Aos meus colegas do curso de pós-graduação em Análise de Dados pela troca de opiniões e ajuda mútua. Juntos, conseguimos avançar e ultrapassar vários obstáculos. Em especial à amiga Sarah Lima Bezerra, companheira em vários trabalhos realizados em grupo e parceira no compartilhamento de ideias.

A todos os meus professores do curso de pós-graduação em Análise de Dados para o Controle, pela excelência da qualidade técnica de cada um.

Aos meus chefes Maurício Ramos e Silva, Eduardo Chaves Ferreira e José Renato Alves Affonso pelo apoio e compreensão da importância dada à esta iniciativa.

E a todos que direta ou indiretamente fizeram parte da minha trajetória, o meu muito obrigado.

# RESUMO

Os hospitais são estabelecimentos de saúde altamente complexos que lidam com grande diversidade de recursos humanos, especialidades médicas, volume de materiais e medicamentos. Enfrentam um enorme desafio para gerenciar suas atividades, com recursos cada vez mais escassos, em meio ao aumento crescente na demanda de seus serviços. No Brasil, os gastos federais com SUS são enormes, e a atenção hospitalar é responsável por consumir boa parte desses recursos. O Tribunal de Contas da União, como órgão de Controle Externo do governo federal, é responsável por avaliar a eficiência e qualidade dos serviços públicos prestados pelo governo federal, entre eles, a eficiência na atenção hospitalar. Mas como realizar esse diagnóstico tendo em vista o grande volume de dados existente sobre esse assunto? É possível o emprego de recursos de TI para tornar mais fácil o alcance dessa missão? É possível o uso de ferramentas de ciência de dados para ajudar de alguma forma essa tarefa? O presente trabalho se enquadra nesse contexto. Ele propõe o uso de técnicas de aprendizagem de máquina na tentativa de selecionar, dentre os hospitais cadastrados no Sistema CNES, aqueles com perfis de atendimento mais similares. A metodologia utilizada durante a realização do trabalho foi a *CRISP-DM*, e os algoritmos de mineração aplicados se enquadram na categoria dos não supervisionados de clusterização. Os resultados obtidos foram analisados e comparados de modo a constatar sua eficácia no reconhecimento de padrões por vias computacionais, distinguindo perfis de hospitais por tipo de atendimento prestado.

**Palavras-chave:** Tribunal de Contas da União. SUS. CNES. Eficiência. *CRISP-DM*. Clusterização.

# ABSTRACT

*Hospitals are high complex healthcare units that deals with a wide range of human resources, medical specialties, materials and drugs. They face an enormous challenge for managing their activities, with limited budget, among the increased demand for their services. In Brazil, federal spending on SUS (Brazilian Healthcare System) is huge, and hospital is responsible for consuming a large part of these resources. The Federal Court of Accounts, in the role of external control body, is responsible for evaluating public services quality and efficiency performed by federal government, including hospital care programs. But how to perform this diagnosis towards the gigantic volume of data available about this subject? Is it possible to make use of IT resources to facilitate this mission? Is it possible to use data science tools to help this task? The present work fits this context. It proposes the use of machine learning techniques to attempt gathering, among the hospitals registered at CNES system, those with most similar care profiles. The methodology used during this project was CRISP-DM, and the data mining algorithms applied belongs to unsupervised clustering category. The results reached were analyzed and compared in order to verify its recognition patterns effectiveness using computational resources, distinguishing hospital by care profiles types.*

***Keywords:*** *Federal Court of Accounts. Brazilian Healthcare System. Efficiency. CRISP-DM. Clustering*

## LISTA DE ILUSTRAÇÕES

## LISTA DE TABELAS

# LISTA DE ABREVIATURAS E SIGLAS

| | |
|---|---|
| CF88 | Constituição da República Federativa do Brasil de 1988 |
| CNES | Cadastro Nacional de Estabelecimentos de Saúde |
| CRS | *Constant Returns to Scale* |
| DASHBOARD | Painéis de visualização de dados |
| DATA MIMING | Mineração de dados |
| DEA | *Data Envelopment Analysis* |
| DMU | *Decision Making Unit* |
| LABCONTAS | Repositório de dados do datacenter do TCU |
| MER | Modelo de Entidade Relacionamento |
| MS | Ministério da Saúde |
| OCDE | Organização para o Comércio e Desenvolvimento |
| SETIC | Secretaria de Infraestrutura de Tecnologia da Informação |
| SIA | Sistema de Informação Ambulatorial |
| SIH | Sistema de Informação Hospitalar |
| SUS | Sistema Único de Saúde |
| TCU | Tribunal de Contas da União |
| TI | Tecnologia da Informação |
| UFRJ | Universidade Federal do Rio de Janeiro |

# SUMÁRIO

# 1 INTRODUÇÃO

A Constituição Federal de 1988 - CF88, no Art. 196, inovou ao declarar que “Saúde é direito de todos e dever do Estado” (BRASIL, 1988). No período anterior à CF88, esse entendimento não era assim. O sistema público de saúde pretérito possibilitava prestação e assistência à saúde apenas aos trabalhadores vinculados à Previdência Social. Aos demais cidadãos, aqueles que não possuíam vínculo oficial com a previdência, restava apenas o atendimento pelas entidades filantrópicas.

A criação do Sistema Único de Saúde – SUS ocorreu logo após a promulgação constitucional, por intermédio da Lei nº. 8.080 (BRASIL, 1990). Esta Lei possibilitou o acesso ao sistema público de saúde de forma universal e sem discriminação. Atualmente, a atenção de saúde é integral, abrangendo não somente os cuidados assistenciais, como também a prevenção e a promoção da saúde, desde a gestação e por toda a vida, com foco na saúde com qualidade de vida.

O SUS é um dos maiores e mais complexos sistemas de saúde pública do mundo. Ele abrange desde um simples atendimento para avaliação da pressão arterial, por meio da atenção primária, até procedimentos extremamente complexos, como transplante de órgãos (SAÚDE, 2020). Com a sua criação, o SUS operacionalizou o acesso universal ao sistema público de saúde definido na Constituição.

A CF88 (BRASIL, 1988) também inovou ao definir que a gestão das ações e dos serviços de saúde deve ser solidária e participativa entre os três entes da Federação: a União, os Estados e os Municípios, cabendo a cada Ente Federativo seu quinhão de responsabilidade e competência. A rede que compõe o SUS atualmente é ampla e abrange desde os serviços básicos até as internações de média e alta complexidades, incluindo serviços de emergências, vigilância epidemiológica, sanitária, ambiental e assistência farmacêutica.

A operação do SUS está sob a tutela do Ministério da Saúde, o qual, a fim de cumprir a Lei de Acesso à Informação (BRASIL, 2011), disponibiliza os dados abertos para consulta aos cidadãos por intermédio do site www.datasus.gov.br/DATASUS.

Um dos sistemas que operacionaliza o SUS é o CNES – Cadastro Nacional de Estabelecimentos de Saúde, criado para unificar e organizar os dados sobre unidades de saúde no Brasil (SAÚDE, 2019). O sistema CNES reúne informações sobre as equipes, profissionais e infraestrutura, tais como leitos disponíveis, equipamentos utilizados etc. Através do CNES, o Ministério da Saúde toma ciência dos consultórios, clínicas e hospitais presentes nas cidades

brasileiras. Os dados permitem o gerenciamento dos serviços de saúde disponíveis para a população, servindo como base, por exemplo, para a avaliação dos locais que precisam de mais leitos hospitalares.

A concepção de um Sistema de Estabelecimentos de Saúde ocorreu no ano de 2000 com a publicação da Portaria MS/SAS 376 (SAÚDE, 2000), de 4 de outubro. Com a incorporação das sugestões recebidas dos gestores estaduais, municipais e da sociedade em geral, editou-se a Portaria MS/SAS 511 (SAÚDE, 2000), de 29 de dezembro de 2000, que passou a normatizar o processo de cadastramento e recadastramento dos estabelecimentos de saúde em todo o território nacional.

Alguns anos depois, uma auditoria do Tribunal de Contas da União - TCU realizada na Secretaria de Estado da Saúde de Roraima (Sesau/RR) e no Núcleo Estadual do Ministério da Saúde no Estado de Roraima (NEMS/RR), no período compreendido entre 11/8/2014 e 3/10/2014, menciona o sistema CNES, destacando:

> *"O sistema representa um desejo há muito aspirado por todos que utilizam as informações de saúde como base para elaboração do seu trabalho, tanto no aspecto operacional quanto gerencial, visto que os dados cadastrais constituem pontos fundamentais para a elaboração da programação, controle e avaliação da assistência hospitalar e ambulatorial no país, assim como a garantia da correspondência entre a capacidade operacional das entidades vinculadas ao SUS e o pagamento pelos serviços prestados." (TCU, 2014)*

O órgão governamental ainda ressalta o CNES, da seguinte forma:

> *É um gigantesco empreendimento no sentido de adquirir o conhecimento efetivo de como está formado o universo de estabelecimentos que cuidam da saúde da população, desde os grandes centros, até as mais longínquas localidades, tornando visível esse cenário a toda sociedade, e fortalecendo o controle social. A base de dados encontra-se disponível a todos os interessados, podendo ser acessada pela Internet por meio do link http://cnes.datasus.gov.br/. (TCU, 2014)*

# 2 CONTEXTO

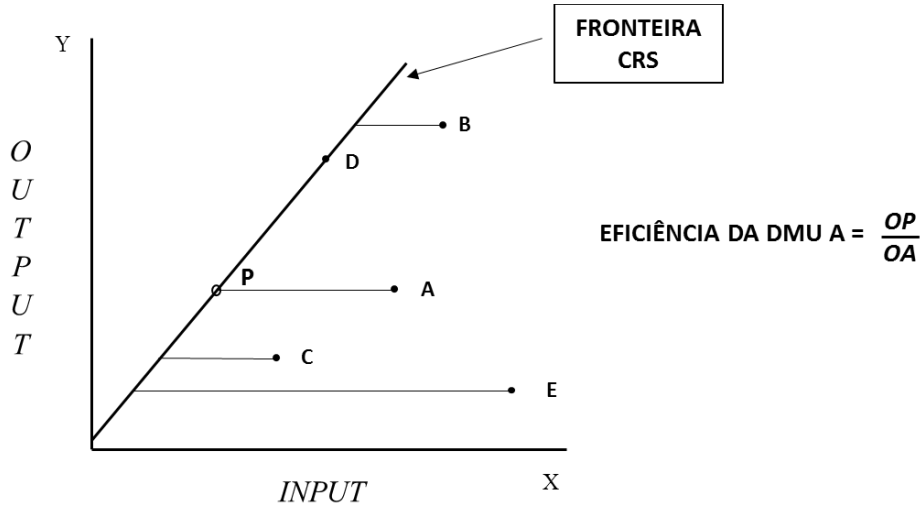
No mundo todo, as despesas com saúde pública têm crescido de forma assustadora a cada ano. Devido a transição demográfica ainda em curso em alguns países, bem como o crescente envelhecimento da população, os gastos com saúde têm tomado proporções surpreendentes. No Brasil isso não é diferente. O Total de despesas destinadas ao Sistema Único de Saúde - SUS em 2019 ultrapassou a casa do R$ 114 bilhões (FNS, 2019). Deste montante, 67% destinam apenas para a atenção hospitalar, percentual bem superior à média observada entre os países da Organização para o Comércio e Desenvolvimento - OCDE, onde os gastos cuidados hospitalares representam 38% do total gasto em saúde (OCDE, 2020).

Um dos projetos da SecexSaúde, Secretaria do Tribunal de Contas da União, previstos para o biênio 2019 / 2020, é "realizar o levantamento da eficiência das unidades prestadoras de serviço de saúde de média e alta complexidades" (TCU, 2019). Este levantamento, ora em curso, utiliza recursos de ciência de dados, mais especificamente o emprego da técnica DEA (*Data Envelopment Analysis*).

DEA é uma técnica econométrica de análise de eficiência muito utilizada em pesquisas operacionais, nas quais as unidades de saúde se enquadram perfeitamente. Esta técnica é utilizada para estimar a eficiência relativa de unidades, denominadas DMU (*Decision Making Unit*), comparando a produção de cada uma delas. Consiste em confrontar unidades (DMU) que produzem mesmas saídas (*outputs*) com mesmas entradas (*inputs*). No caso do trabalho em questão, DMUs são unidades hospitalares cadastradas no sistema CNES. As entradas são as variáveis que refletem os recursos disponíveis na unidade, como médicos, enfermeiros, leitos, equipamentos etc, e as saídas são os atendimentos, consultas e internações realizadas. A eficiência é calculada dividindo as saídas pelas entradas, e comparando seus resultados. Teoricamente, quem produz mais com menos é mais eficiente.

De forma gráfica, desenhando as eficiências das unidades, pode-se observar que as mais eficientes ficam na fronteira do gráfico, enquanto que as demais (abaixo da linha), são menos eficientes.

Figura 1 - Curva de eficiência relativa da análise DEA




Fonte: (GONÇALVES e NORONHA, 2002)

Na Figura 1 - Curva de eficiência relativa da análise DEA, os pontos A, B, C e E correspondem às unidades ineficientes. O ponto D é a unidade eficiente, que está sobre a reta, e representa a fronteira eficiente, CRS (*Constant Returns to Scale*). O deslocamento de A para a fronteira eficiente (ponto P) implicaria o valor ótimo do input que tornaria esta unidade eficiente (GONÇALVES e NORONHA, 2002)

Já existe uma farta literatura que descreve a utilização de métodos DEA em unidades de saúde no mundo inteiro. No Brasil, vários trabalhos publicados utilizaram essa técnica, como por exemplo o trabalho sobre a aplicação da análise envoltória de dados para avaliar a eficiência de hospitais do SUS em Mato Grosso (SOUZA, SCATENA e KEHRIG, 2016). Outro trabalho muito interessante, do Banco Mundial em parceria com a UFRJ, abordou este problema realizando análises separadamente para hospitais gerais e hospitais especializados (BM, 2019).

Embora essa técnica seja bastante utilizada, há diversas críticas em relação a seu emprego, principalmente no que tange à seleção das DMUs comparáveis. De fato, os resultados da análise DEA pode revelar resultados equivocados, caso as DMUs escolhidas não sejam essencialmente homogêneas. Hospitais gerais, por exemplo, não constituem um grupo totalmente homogêneo, haja visto haver muita heterogeneidade entre ele. Compará-los sem nenhum critério de seleção, parece equivocado.

Mas como garantir uma homogeneidade nas DMUs comparáveis? Como certificar que as unidades são essencialmente similares? É nesse contexto que se apresenta esse trabalho. A utilização de recursos tecnológicos disponibilizados atualmente no TCU, bem como a aplicação de ferramentas da área de ciência de dados, torna possível o emprego de técnicas de mineração

que sejam capazes de selecionar DMUs homogêneas entre si, visando a aplicação da técnica DEA e viabilizar a comparação entre unidades realmente similares, podendo ser bastante relevante para melhorar o resultado do trabalho da SecexSaúde.

Este documento está organizado em forma de seções. A seção 3 apresenta os problemas identificados para os quais se pretende propor as soluções objeto deste trabalho, bem como a justificação para sua realização. A seção 4 descreve a metodologia utilizada, a qual norteou a execução das atividades executada durante o trabalho. As seções 5 à 10 narram as etapas da metodologia utilizada, iniciando pelo entendimento do negócio (seção 5), passando por entendimento dos dados (seção 6), preparação dos dados (seção 7), modelagem (seção 8), avaliação (seção 9) e finalizando na aplicação (seção 10). Por fim, a seção 11 apresenta as considerações finais sobre o assunto.

## 3 PROBLEMA E JUSTIFICATIVA

Em entrevista com a SecexSaúde, foi constatado a existência de um projeto em curso na unidade, como meta estratégica do biênio 2019 / 2020, de levantamento das unidades de saúde de média e alta complexidades (TCU, 2019). Este projeto teria como objetivo realizar diagnóstico da eficiência na gestão de saúde dos hospitais do Brasil com aplicação da DEA que, por meio do uso de técnicas de programação linear, pode ser empregada na aferição da eficiência relativa de unidades hospitalares. Foi constatado também que o TCU, até então, não possuía internalizado no seu datacenter todos os dados necessários para esta análise, que são os dados do sistema CNES – Cadastro Nacional de Estabelecimentos de Saúde, SIH – Sistema de Internações Hospitalares e SIA - Sistema de Atendimentos Ambulatoriais.

A ausência destes dados no ambiente operacional de TI do Tribunal acarretava diversos inconvenientes operacionais, como a dificuldade na realização de consultas em grandes volumes de dados sobre as unidades, bem como a impossibilidade de cruzamento deles com outras bases de dados já internalizadas na Corte de Contas. Além disso, a ausência desses dados causava a impossibilidade de aplicação da técnica DEA, comprometendo o sucesso do projeto da Unidade de Negócio.

Diante desses fatos, chegou-se à constatação dos seguintes problemas:

1) Necessidade de internalizar os dados dos sistemas CNES no datacenter do TCU;
2) Agrupar unidades de saúde do (DMUs) em subconjuntos homogêneos de forma a tornar mais realista o resultado da aplicação da análise DEA.

Por fim, este trabalho se justifica, pois, um dos objetivos estratégicos do Tribunal de Contas da União, definidos no Plano Estratégico 2015-2021 (TCU, 2015), é “aprimorar o uso da TI como instrumento de inovação para o controle”. O emprego apropriado dos recursos de TI torna mais ágil e focada a atuação do Tribunal no alcance de sua missão, assim como amplia o universo de recursos fiscalizados e a capacidade de resposta às demandas apresentadas. Possibilitar o uso da TI nos trabalhos da SecexSaúde, certamente melhorará a assertividade dos achados de auditorias e fiscalizações realizadas por esta Unidade Técnica.

# 4 METODOLOGIA

Este trabalho foi desenvolvido utilizando a Metodologia *CRISP-DM* (*Cross Industry Standart Process for Data Mining*), amplamente utilizada em projetos relacionados a mineração de dados (IBM, 2014). Surgiu em 1996 a partir da experiência de três empresas pioneiras no setor: DaimlerChrysler, que aplica análises de mineração em seus negócios, NCR, que provê soluções de *datawarehouse* e SPSS, que disponibiliza soluções baseadas no processo de mineração de dados.

Essa metodologia proporciona a padronização dos passos de um projeto de análise de dados. Sua aplicação pode ser usada em diferentes mercados, independente do segmento, possibilitando também que projetos de análise de dados sejam concluídos mais rapidamente, mais eficientemente e com um custo menor.

O *CRISP-DM* é dividido em seis etapas. A seguir, uma breve exposição das etapas do *CRISP-DM*:

Tabela 1 - Etapas do *CRISP-DM*

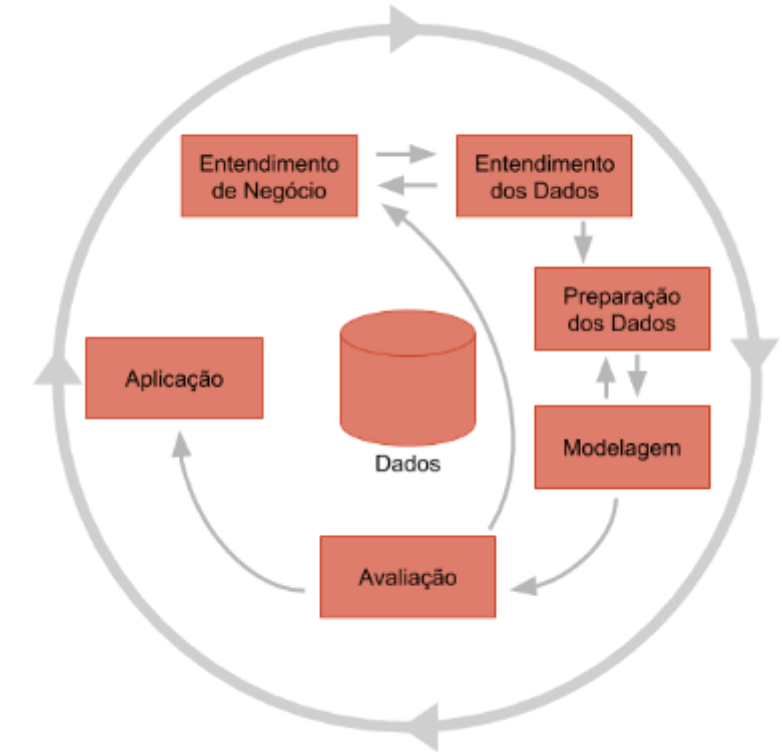
| FASE | OBJETIVO |
|---|---|
| Entendimento do Negócio | Foca em entender o objetivo do projeto a partir de uma perspectiva de negócios, definindo um plano preliminar para atingir os objetivos |
| Entendimento dos Dados | Recolhimento de dados e início de atividades para familiarização com os dados, identificando problemas e metas |
| Preparação dos Dados | Construção do conjunto de dados final, a partir dos dados iniciais. Normalmente ocorre várias vezes no processo |
| Modelagem | Onde várias técnicas de modelagem são aplicadas, e seus parâmetros calibrados para otimização. Assim, é comum retornar à Preparação dos Dados durante essa fase |
| Avaliação | Onde o modelo construído é testado para validar se atendem às necessidades do negócio. |
| Aplicação | O conhecimento adquirido pelo modelo é organizado e apresentado de uma maneira que o cliente possa utilizar |

Fonte: Elaborada pelo autor (2020)

No decorrer deste trabalho, cada fase da metodologia descrita na Tabela 1 - Etapas do *CRISP-DM*, acima, foi detalhada.

A Figura 2 - Diagrama de iteração entre fases do *CRISP-DM*, abaixo, mostra a sequência mais comum e frequente entre as fases do *CRISP-DM*. O círculo externo no diagrama representa a natureza cíclica do próprio projeto de análise de dados, simbolizando que a mineração continua com diversas iterações após a implantação da solução.

Figura 2 - Diagrama de iteração entre fases do *CRISP-DM*



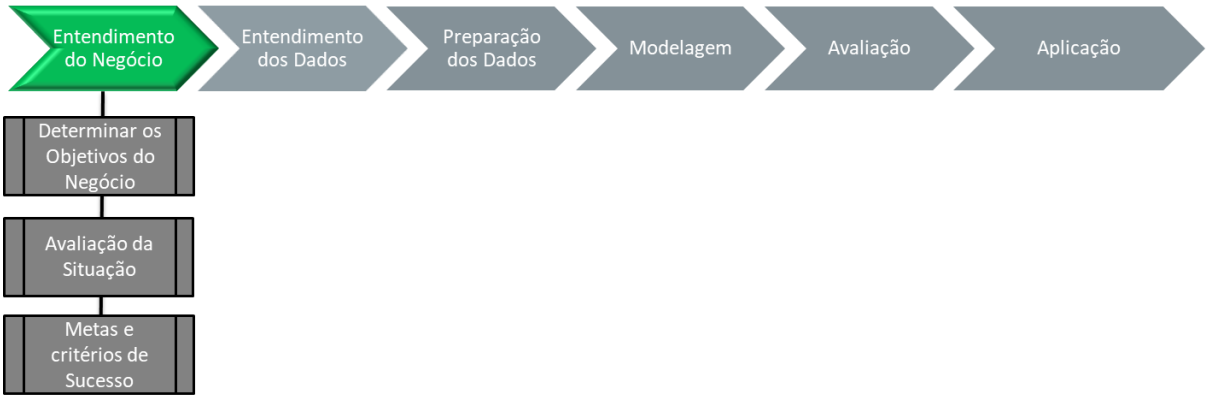

Fonte: (MEDIUM, 2020)

# 5 ENTENDIMENTO DO NEGÓCIO

Entendimento do Negócio é a fase do *CRISP-DM* que visa explorar o que se espera obter da mineração de dados. Nela, desenvolvem-se atividades para entender o negócio do ponto de vista do cliente, envolvendo diversas pessoas chave, conhecedoras do negócio. Embora pareça dispensável, esta fase é crucial para todo trabalho, pois é a partir dela que se avaliam os motivos do negócio, bem como os esforços necessários para a mineração de dados (IBM, 2014).

A Figura 3 - Atividades da fase "Entendimento do negócio" do *CRISP-DM*, abaixo, mostra as atividades realizadas nesta etapa.

Figura 3 - Atividades da fase "Entendimento do negócio" do *CRISP-DM*




Fonte: Elaborada pelo autor (2020)

## 5.1 DETERMINAR OS OBJETIVOS DE NEGÓCIO

Dado o contexto exposto, e uma vez definidos os problemas que se propõem resolver neste trabalho, determinaram-se os objetivos geral e específicos que se desejam alcançar, detalhados a seguir.

### 5.1.1 Objetivo Geral

Este trabalho tem como objetivo geral a clusterização das unidades hospitalares cadastradas no CNES, com vistas à criação de grupos de unidades homogêneas para subsidiar a implementação da DEA no contexto do projeto de Levantamento da Eficiência das Unidades Prestadoras de Serviço de Saúde de Média e Alta Complexidades (TCU, 2019).

### 5.1.2 Objetivos Específicos

Para alcançar o objetivo geral, vislumbrou-se a necessidade de executar outras atividades, derivando nos seguintes objetivos específicos:

a) Internalizar os dados do Sistema CNES no Labcontas (repositório de dados do TCU), tornando possível a realização de consultas em grande volume de dados e de forma célere. Esta ação possibilitará também a realização de cruzamento de dados com outras bases de dados já internalizadas na Corte, até então não possíveis, abrindo um leque enorme de novas possibilidades de análises e auditorias.
b) Explorar as diversas alternativas de clusterização nas unidades de saúde do CNES de forma a identificar e selecionar a melhor opção a ser empregada na técnica DEA.
c) Criar forma de visualização dos dados do sistema CNES, como também visualização dos grupos formados na clusterização. Esta ação permitirá a análise e manipulação dos resultados, no ambiente do TCU, e de forma mais fácil, rápida e amigável pelos auditores e analistas da Secretaria.

## 5.2 AVALIANDO A SITUAÇÃO

O sistema CNES é um dos principais sistemas do SUS, e serve como fonte de informação primária para consulta de unidades de saúde no Brasil. Além disso, ele também é usado para ratificar os dados de atendimentos hospitalares e ambulatoriais registrados nos sistemas SIA e SIH.

Um dos objetivos principais do sistema CNES, conforme Portaria No 1.646 de 2015 (SAÚDE, 2015), é de automatizar o processo de coleta de dados feita nos Estados e Municípios sobre a capacidade física instalada das unidades de saúde, informando quais os serviços disponíveis e profissionais vinculados aos estabelecimentos de saúde, equipes alocadas, subsidiando os gestores com dados de abrangência nacional para efeito de planejamento de ações em saúde (SAÚDE, 2015). Os critérios mínimos para se considerar uma edificação como um estabelecimento de saúde são:

- Possuir espaço físico delimitado e permanente, isto é, possuir uma infraestrutura necessária para realização dos atendimentos de saúde. Podem estar incluídos estabelecimentos móveis como embarcações, carretas, etc. Entretanto, espaços temporários estão excluídos, como barracas, tendas ou atendimentos abertos em locais públicos.
- Realizar atendimentos de saúde humana, ou seja, é imprescindível que haja a obrigatoriedade do efetivo funcionamento. Espaços ainda em construção, não abertos ou desativados não podem ser considerados como um estabelecimento de saúde. Importante salientar que saúde humana abrange não somente a prestação

de caráter assistencial, como também os estabelecimentos que realizam ações de vigilância, regulação ou gestão da saúde. Estabelecimentos que não têm o foco direto na saúde humana, como os que visam a saúde animal, os salões de beleza e as clínicas de estética não devem ser considerados como estabelecimentos de saúde.

- Possuir responsabilidade técnica, isto é, os serviços prestados devem estar regidos por legislação, associando o estabelecimento de saúde à um profissional de saúde (pessoa física) legalmente responsável por elas.

## 5.3 METAS DE MINERAÇÃO E CRITÉRIOS DE SUCESSO

Para cada um dos objetivos litados no item 5.1 DETERMINAR OS OBJETIVOS DE NEGÓCIO, derivou-se uma meta de mineração correspondente, que por sua vez resultou em um ou mais critérios de sucesso. As premissas adotadas neste trabalho levaram em consideração que as metas de mineração seriam atingidas caso os critérios de sucesso fossem satisfeitos. Na Tabela 2 – Metas de mineração e critérios de sucesso, abaixo, estão listadas as metas com seus respectivos critérios.

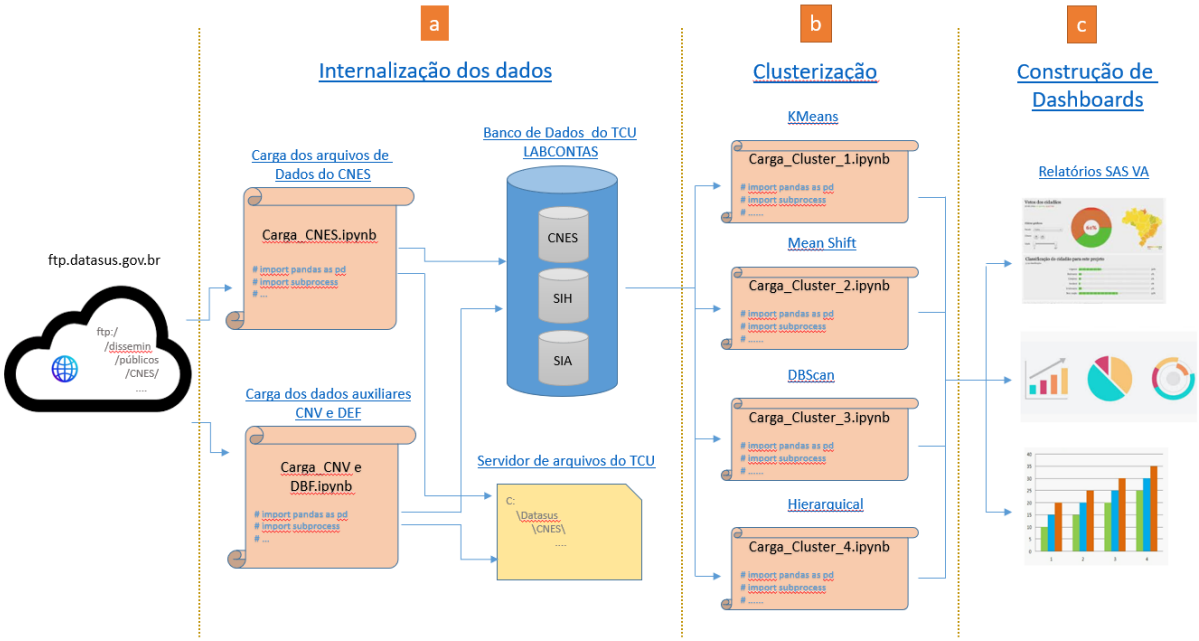
Tabela 2 – Metas de mineração e critérios de sucesso

| # | META DE MINERAÇÃO | CRITÉRIO DE SUCESSO |
|---|---|---|
| a | Construir algoritmo para importar os dados do sistema CNES, originários do site do DATASUS, e internalizá-los no repositório de dados do TCU (Labcontas). | - O algoritmo deve ser capaz de realizar a carga inicial dos dados do sistema CNES, bem como a atualização dos dados incrementais, na medida em que novos arquivos são inseridos no site do DATASUS<br>- O algoritmo deve utilizar linguagem *Python*, empregando os conhecimentos adquiridos durante o curso de Análise de Dados para o Controle |
| b | Construir modelo de mineração capaz de agrupar unidades de saúde (DMU) em subconjuntos homogêneos de unidades de saúde | - Modelo deve empregar uma das técnicas de mineração de dados aprendidas durante o curso de Análise de Dados.<br>- Deve-se utilizar mais de um algoritmo, de forma a possibilitar a comparação e análise de seus resultados.<br>- Deve ser capaz de agrupar unidades de saúde com perfil similares e homogêneos de atendimentos. |
| c | Construir painéis para visualização dos dados | - Deve ser capaz de demonstrar os resultados obtidos nas análises de forma rápida, amigável e manipulável pelo usuário.<br>- Deve ser construído utilizando ferramenta homologada pela Secretaria de Infraestrutura de Tecnologia da Informação do TCU – SETIC |

Fonte: Elaborada pelo autor (2020)

A Figura 4 – Produtos desenvolvidos no trabalho, abaixo, mostra os artefatos produzidos durante o trabalho, que englobam todas as três metas de mineração definidas na Tabela 2 – Metas de mineração e critérios de sucesso.

Figura 4 – Produtos desenvolvidos no trabalho




Fonte: Elaborada pelo autor (2020)

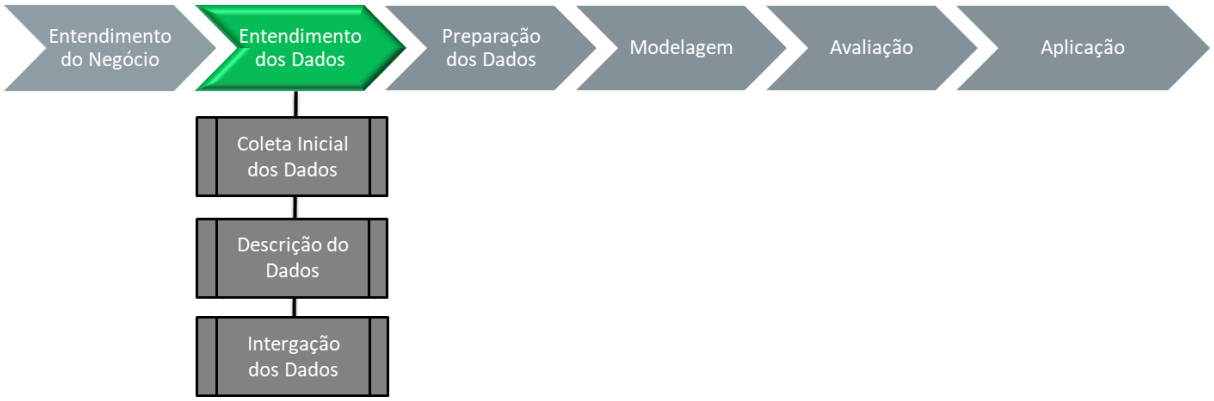
# 6 ENTENDIMENTO DOS DADOS

A fase de Entendimento de Dados definida pelo *CRISP-DM* visa apreciar com maior detalhe os elementos disponíveis para a realização da mineração. Envolve acessar os dados, explorá-los e realizar análise de qualidade neles. Pode envolver também a geração de gráficos que permitam melhorar a compreensão dos mesmos.

Essa etapa é fundamental para evitar problemas inesperados durante todas as demais etapas seguintes, principalmente a etapa de Preparação de Dados, que é normalmente a parte mais longa de um projeto (IBM, 2014).

A Figura 5 - Detalhamento da fase "Entendimento dos Dados" do *CRISP-DM*, abaixo, mostra as atividades realizadas nesta etapa.

Figura 5 - Detalhamento da fase "Entendimento dos Dados" do *CRISP-DM*




Fonte: Elaborada pelo autor (2020)

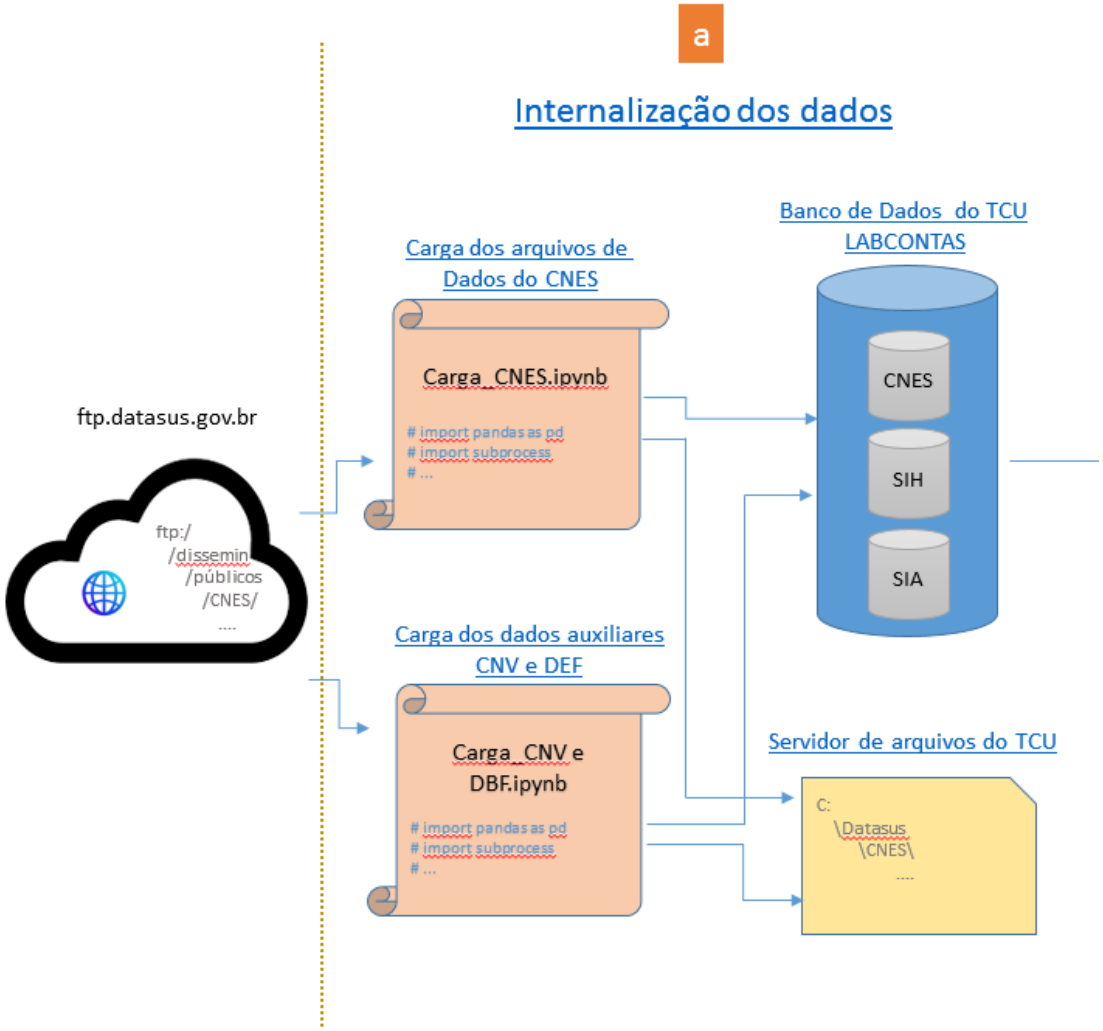
## 6.1 COLETA INICIAL DOS DADOS

A primeira parte do trabalho focou na coleta dos dados necessários para a realização deste trabalho. Tendo em vista que os dados do sistema CNES **não** estavam disponíveis no datacenter do TCU, definiu-se por construir mecanismo para a internalização deles.

Embora existisse algum dado disponível do SIA e SIH no datacenter do TCU, os mesmos estavam desatualizados. Optou-se, porém, por **não** realizar atividades alguma de carga desses dados, tendo em vista que outros trabalhos de conclusão de curso, em andamento no TCU, e executados em paralelo a este, tinham como metas a realização destas cargas.

Assim sendo, a coleta inicial dos dados contemplou a meta (a) de mineração definida na Figura 6 – Internalização dos dados, abaixo.

Figura 6 – Internalização dos dados




Fonte: Elaborada pelo autor (2020)

Como pode ser visto na Figura 6 – Internalização dos dados, acima, esta etapa contemplou duas atividades distintas de carga: Carga dos dados auxiliares (CNV e DBF) e Carga dos arquivos de dados do sistema CNES. Ambas as cargas foram desenvolvidas utilizando a linguagem de programação *Python* (PYTHON, 2001), e consistiram em acessar o site do DATASUS, realizar o download dos arquivos para o servidor de arquivos do TCU e efetuar a carga desses dados no repositório de dados do TCU (Labcontas).

### 6.1.1 Carga dos dados auxiliares (CNV e DBF)

Dados auxiliares são dados que contém informações estáticas usadas pelas tabelas de dados do sistema CNES. São arquivos que contém informação de definição (dbf) e conversão (cnv) para efetuar tabulações sobre as bases de dados distribuídas pelo DATASUS. Exemplos

de dados auxiliares são: “tipo de unidade”, “turno de atendimento”, “nome fantasia do cnes” etc. Tendo em vista sua característica estática, optou-se por internalizá-los uma única vez no TCU. Os arquivos originais encontram-se encapsulados no arquivo TAB_CNES.zip no caminho: ftp://ftp.DATASUS.gov.br/dissemin/publicos/CNES/200508_/Auxiliar/ (DATASUS, 2017).

Os dados auxiliares utilizados neste trabalho foram carregados no Labcontas no esquema BDU_SECEXSAUDE_CNES.dbo. A título de exemplo, alguns estão listados na Tabela 3 – Exemplos de arquivo CNV internalizados, abaixo.

Tabela 3 – Exemplos de arquivo CNV internalizados

| TABELA | DESCRIÇÃO |
|---|---|
| TAB_AUXILIAR_ATIVIDADE | Contém informações sobre a atividade de ensino dos estabelecimentos |
| TAB_AUXILIAR_CBO | Contém informações sobre o Código Brasileiro de Ocupação utilizadas nos estabelecimentos |
| TAB_AUXILIAR_EQUIPAMENTOS | Contém os tipos de equipamentos existentes nos estabelecimentos |
| TAB_AUXILIAR_ESTABELECIMENTOS | Contém os nomes e endereços dos estabelecimentos |
| TAB_AUXILIAR_NATJUR | Contém as naturezas jurídicas dos estabelecimentos |
| TAB_AUXILIAR_TURNO | Contém os turnos praticados nos estabelecimentos |
| TAB_AUXILIAR_TP_ESTABELECIMENTOS | Contém os tipos de estabelecimentos |
| TAB_AUXILIAR_TP_LEITOS | Contém os tipos de leitos dos estabelecimentos |
| TB_SUBGR | Contém os nomes dos procedimentos utilizados a nível de SUBGRUPO do código SIGTAP |

Fonte: Elaborada pelo autor (2020)

### 6.1.2 Carga do CNES (arquivos DBC)

Consistem nos dados dinâmicos das unidades de saúde do Brasil. São atualizados de forma incremental pelo Ministério da Saúde, e possuem a extensão dbc. Estes são os dados que possuem a “informação” propriamente dita do sistema CNES, os quais foram efetivamente usados na clusterização. Possuem periodicidade mensal e estão disponíveis no site do DATASUS desde o mês de agosto de 2005. Por questões de economia de espaço em disco, bem como tempo de execução, optou-se por internalizar os dados apenas a partir do mês de janeiro de 2012.

Para que fosse possível a leitura dos arquivos de extensão dbc pela linguagem de programação *Python* (PYTHON, 2001), foi utilizada a biblioteca *pysus* (COELHO, 2018). Durante a fase de execução deste algoritmo, foi observado alto tempo de processamento na carga para o repositório de dados do TCU (Labcontas). Para contornar esse problema, optou-se pela execução do comando *bulk insert*, cuja função principal consiste em carregar lotes de

registros direto no banco. Para isso, foi necessário converter arquivos dbc para csv, os quais foram armazenados no servidor de arquivos do TCU.

Os arquivos originais foram obtidos do DATASUS do caminho: ftp://ftp.DATASUS.gov.br/dissemin/publicos/CNES/200508_/Dados/, e a descrição dos campos é feita nos itens que seguem.

## 6.2 DESCRIÇÃO DOS DADOS

Os dados utilizados neste trabalho foram obtidos dos sistemas CNES, SIH e SIA, sendo que apenas para o CNES foi realizada a internalização no repositório do TCU. Os demais foram obtidos diretamente do Labcontas. A seguir, estão detalhados os campos utilizados em cada um deles.

### 6.2.1 Dados do sistema CNES

Os dados do sistema CNES são os provenientes dos arquivos dbc, e estão divididos em treze tabelas diferentes, cada qual contendo um determinado tipo de informação sobre a unidade de saúde. Os arquivos estão separados por unidade da federação, e por mês / ano. As nomenclaturas dos arquivos estão padronizadas da seguinte forma: os dois primeiros caracteres contêm o tipo de arquivo; os dois caracteres seguintes contêm a unidade da federação; os dois seguintes contêm o ano, e nos dois últimos, o mês de referência (DATASUS, 2017).

Para manter a padronização dos dados de acordo com a origem, foram criadas treze tabelas no repositório de dados do TCU (Labcontas), no esquema denominado BDU_SECEXSAUDE_CNES.dbo. Para cada tipo de dado, uma tabela específica foi criada. A seguir contém o detalhamento dos arquivos obtidos no site do DATASUS, bem como as tabelas criadas no Labcontas.

Tabela 4 - Arquivos do sistema CNES

| | TIPO DE DADO | NOME DO ARQUIVO NA ORIGEM | NOME DA TABELA NO DESTINO |
|---|---|---|---|
| 1 | Estabelecimentos | STufaamm.dbc | ST |
| 2 | Dados Complementares | DCufaamm.dbc | DC |
| 3 | Profissional | PFufaamm.dbc | PF |
| 4 | Leitos | LTufaamm.dbc | LT |
| 5 | Equipamentos | EQufaamm.dbc | EQ |
| 6 | Serviço Especializado | SRufaamm.dbc | SR |
| 7 | Equipes | EPufaamm.dbc | EP |
| 8 | Habilitações | HBufaamm.dbc | HB |
| 9 | Regras Contratuais | RCufaamm.dbc | RC |
| 10 | Gestão e Metas | GMufaamm.dbc | GM |
| 11 | Estabelecimento de Ensino | EEufaamm.dbc | EE |
| 12 | Estabelecimento Filantrópico | EFufaamm.dbc | EF |

| | TIPO DE DADO | NOME DO ARQUIVO NA ORIGEM | NOME DA TABELA NO DESTINO |
|---|---|---|---|
| 13 | IntegraSUS | INufaamm.dbc | IN |

Fonte: (DATASUS, 2017)

Conforme já mencionado, os dados do sistema CNES são atualizados mensalmente no site do DATASUS. A quantidade de registros de cada arquivo varia em cada mês, a medida em que unidades de saúde são credenciadas ou descredenciadas do sistema. Para fins de exemplo, a Tabela 5 - Quantitativo de registros do sistema CNES de dezembro de 2018, a seguir, contém a lista das tabelas e seus respectivos registros obtidos neste mês em questão.

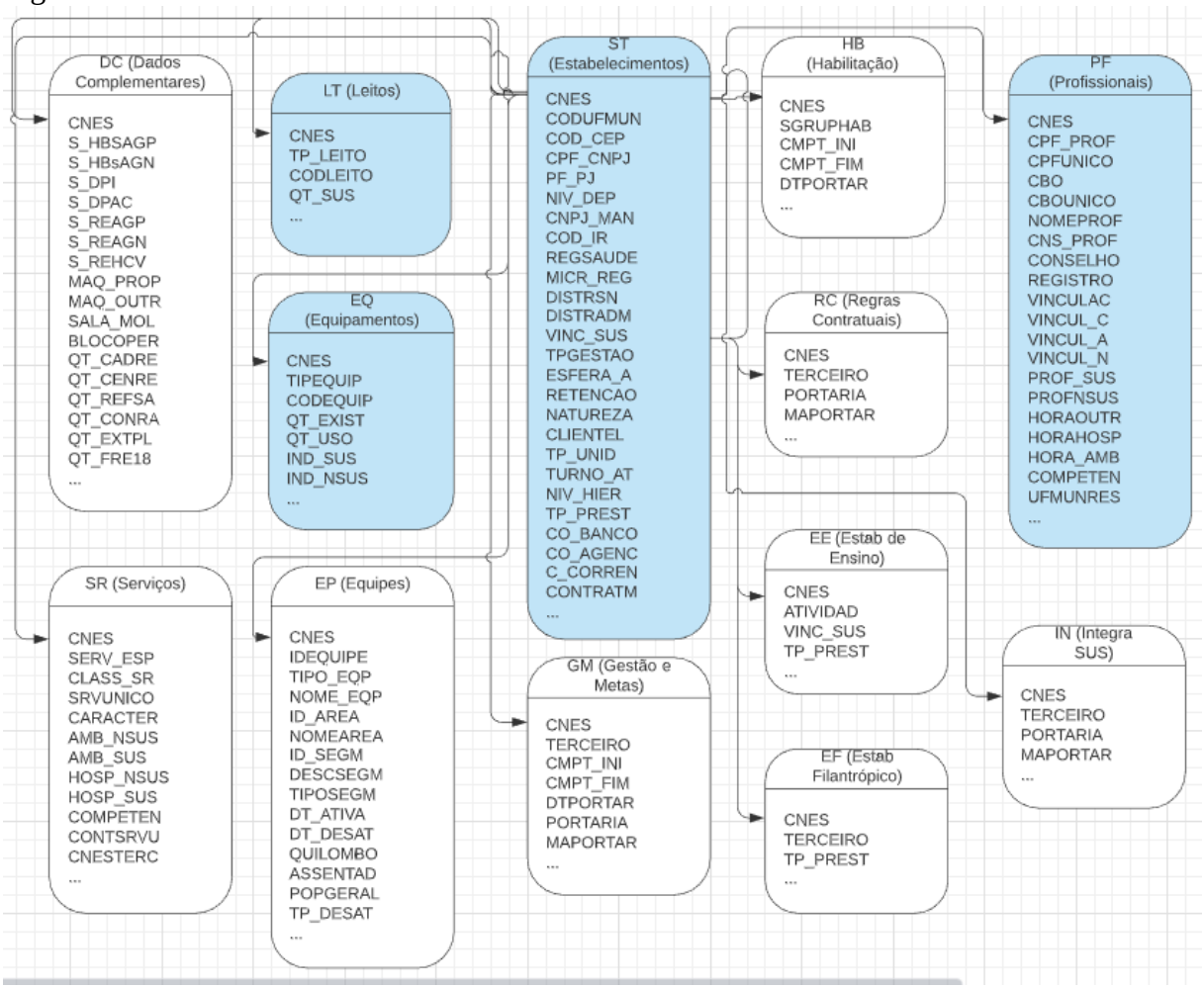
Tabela 5 - Quantitativo de registros do sistema CNES de dezembro de 2018

| | TABELA | QTD DE REGISTROS NO MÊS DE DEZEMBRO DE 2018 |
|---|---|---|
| 1 | ST | 331.058 |
| 2 | DC | 3.021 |
| 3 | PF | 3.999.299 |
| 4 | LT | 49.214 |
| 5 | EQ | 778.889 |
| 6 | SR | 694885 |
| 7 | EP | 55.965 |
| 8 | HB | 21.585 |
| 9 | RC | 11.957 |
| 10 | GM | 644 |
| 11 | EE | 200 |
| 12 | EF | 701 |
| 13 | IN | 6.173 |

Fonte: Elaborada pelo autor (2020)

O Modelo de Entidade e Relacionamento – MER, contendo as tabelas do sistema CNES bem como seus relacionamentos é mostrado na Figura 7, abaixo. Nota-se que todas as tabelas possuem um campo comum denominado “CNES”, que embora tenha o mesmo nome do sistema em questão, não se confunde com ele. Este campo possui como significação o código das unidades de saúde no Brasil, que são únicos para os estabelecimentos de saúde, não se permitindo repetições entre unidades diferentes. Este campo foi utilizado para realizar a junção entre as tabelas. Estão destacados na cor azul, as tabelas do sistema CNES cujos campos foram utilizados, de alguma forma, no processo de mineração dos dados, objeto deste trabalho.

Figura 7 - Modelo Entidade Relacionamento do sistema CNES




Fonte: Elaborada pelo autor (2020)

O dicionário de dados contendo a descrição de todos os campos do sistema CNES, com seus respectivos tipos de dados estão disponíveis para consulta no site do DATASUS no endereço: ftp://ftp.datasus.gov.br/dissemin/publicos/CNES/200508_/doc/.

### 6.2.2 Dados dos Sistemas SIH e SIA

Além do sistema CNES, outros dois sistemas do DATASUS foram usados neste trabalho: SIH - Sistema de informações Hospitalares e SIA - Sistema de Informações Ambulatoriais.

O SIH é o sistema responsável pelo registro das internações hospitalares. Sua carga para o datacenter do TCU (Labcontas) ocorreu paralelamente à execução deste trabalho, em outro projeto de fim de curso, e atualmente já estão disponíveis no esquema denominado BDU_SECEXSAUDE_SIH.dbo. O sistema SIH possui duas tabelas: RD (Internações Hospitalares) e SP (Serviços Profissionais), sendo que as informações necessárias para a realização deste trabalho estão na tabela RD. Esta tabela, contém informações das internações

realizadas nas unidades de saúde, incluindo detalhes de tempo de internações e valores pagos. Estes dados foram úteis na clusterização pois, a partir deles, foi realizado o cálculo da proporção de cada procedimento em relação ao total da unidade de saúde.

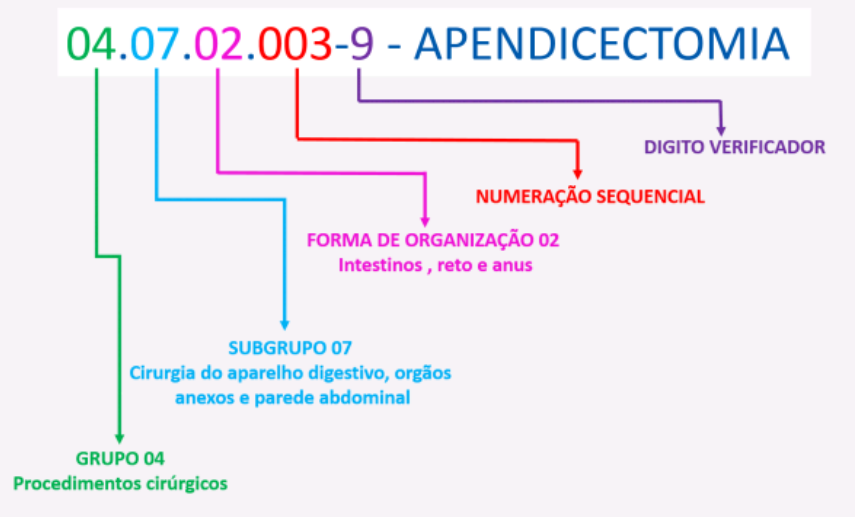
O sistema SIA também pertence ao DATASUS, e contém os procedimentos ambulatoriais realizados nas unidades de saúde do Brasil. Assim como o SIH, a importação dos dados do SIA para o datacenter do TCU ocorreu em paralelo a este trabalho, em outro projeto de fim de curso, e estão atualmente disponíveis no Labcontas em esquema denominado BDU_SECEXSAUDE_SIA.dbo. O SIA possui dez tabelas, sendo que a única utilizada neste trabalho foi a SIA_PA (Procedimentos Ambulatoriais). Esta tabela possui dois tipos de informações: atendimento ambulatorial individual e atendimento ambulatorial consolidado. Para fins deste trabalho, não houve diferenciação quanto aos dois tipos de registros. Os mesmos foram tratados de forma idêntica, respeitando as quantidades e valores pagos. Assim como no SIH, esta informação foi útil para cálculo da proporção de cada procedimento contribuiu para o total realizado no estabelecimento de saúde.

### 6.2.3 Código SIGTAP

Código SIGTAP é forma como o SUS padronizou seus procedimentos. Ele é formado por um conjunto de dez caracteres onde as informações sobre os procedimentos realizados pelo SUS são registrados. Tanto o SIH quanto o SIA utilizam o código SIGTAP em seus atendimentos.

Um exemplo de um código SIGTAP está na figura abaixo, onde o procedimento de "apendicectomia" está traduzido num código SIGTAP.

Figura 8 - Exemplo do código SIGTAP




Fonte: (SIGTAP, 2020)

A informação contida no código SIGTAP obedece ao seguinte formatação: GR.SB.FO.PPP.D, onde: GR corresponde ao grupo a que pertence o procedimento; SB corresponde ao Subgrupo do procedimento no grupo onde está inserido o procedimento; FO corresponde à Forma de organização do procedimento no Subgrupo onde está inserido o procedimento; PPP é o número de ordem sequencial do Procedimento inserido na Forma de organização a qual pertence; por fim o D corresponde ao dígito verificador, que tem como função validar o código do procedimento.

Cabe ressaltar que todo e qualquer procedimento de saúde realizado pelo SUS, tanto no SIA quanto no SIH, requerem necessariamente um código SIGTAP associado a ele. No sistema SIH, o campo que contém o código SIGTAP denomina-se PROC_REA, enquanto que no SIA, o campo denomina-se PA_PROC_ID.

Em entrevista com o cliente, identificou-se que o código SIGTAP deveria ser usado como principal variável na detecção da similaridade de clusterização, tendo em vista que ele contém essencialmente a informação dos atendimentos realizados em cada unidade de saúde.

Ainda em relação ao código SIGTAP, houve ponderação quanto ao nível de agregação a ser usada. Considerando os fatores de tempo de processamento e granularidade dos dados, optou-se por realizar agregação à nível de SUBGRUPO, o que corresponde aos quatro primeiros dígitos. Na Tabela 6 – Códigos SIGTAP agrupados por SUBGRUPO, a seguir, estão listados os tipos de procedimentos por nível de SUBGRUPO realizados pelo SUS, e utilizados neste trabalho.

Tabela 6 – Códigos SIGTAP agrupados por SUBGRUPO

| **CO_SUB_GRU** | **NO_SUB_GRU** |
|---|---|
| 0101 | Ações coletivas/individuais em saúde |
| 0102 | Vigilância em saúde |
| 0201 | Coleta de material |
| 0202 | Diagnóstico em laboratório clínico |
| 0203 | Diagnóstico por anatomia patológica e citopatologia |
| 0204 | Diagnóstico por radiologia |
| 0205 | Diagnóstico por ultrasonografia |
| 0206 | Diagnóstico por tomografia |
| 0207 | Diagnóstico por ressonância magnética |
| 0208 | Diagnóstico por medicina nuclear in vivo |
| 0209 | Diagnóstico por endoscopia |
| 0210 | Diagnóstico por radiologia intervencionista |
| 0211 | Métodos diagnósticos em especialidades |
| 0212 | Diagnóstico e procedimentos especiais em hemoterapia |
| 0213 | Diagnóstico em vigilância epidemiológica e ambiental |

| CO_SUB_GRU | NO_SUB_GRU |
|---|---|
| 0214 | Diagnóstico por teste rápido |
| 0301 | Consultas / Atendimentos / Acompanhamentos |
| 0302 | Fisioterapia |
| 0303 | Tratamentos clínicos (outras especialidades) |
| 0304 | Tratamento em oncologia |
| 0305 | Tratamento em nefrologia |
| 0306 | Hemoterapia |
| 0307 | Tratamentos odontológicos |
| 0308 | Tratamento de lesões, envenenamentos e outros, decorrentes de causas externas |
| 0309 | Terapias especializadas |
| 0310 | Parto e nascimento |
| 0401 | Pequenas cirurgias e cirurgias de pele, tecido subcutâneo e mucosa |
| 0402 | Cirurgia de glândulas endócrinas |
| 0403 | Cirurgia do sistema nervoso central e periférico |
| 0404 | Cirurgia das vias aéreas superiores, da face, da cabeça e do pescoço |
| 0405 | Cirurgia do aparelho da visão |
| 0406 | Cirurgia do aparelho circulatório |
| 0407 | Cirurgia do aparelho digestivo, orgãos anexos e parede abdominal |
| 0408 | Cirurgia do sistema osteomuscular |
| 0409 | Cirurgia do aparelho geniturinário |
| 0410 | Cirurgia de mama |
| 0411 | Cirurgia obstétrica |
| 0412 | Cirurgia torácica |
| 0413 | Cirurgia reparadora |
| 0414 | Bucomaxilofacial |
| 0415 | Outras cirurgias |
| 0416 | Cirurgia em oncologia |
| 0417 | Anestesiologia |
| 0418 | Cirurgia em nefrologia |
| 0501 | Coleta e exames para fins de doação de orgãos, tecidos e células e de transplante |
| 0502 | Avaliação de morte encefálica |
| 0503 | Ações relacionadas à doação de orgãos e tecidos para transplante |
| 0504 | Processamento de tecidos para transplante |
| 0505 | Transplante de orgãos, tecidos e células |
| 0506 | Acompanhamento e intercorrências no pré e pós-transplante |
| 0601 | Medicamentos de dispensação excepcional |
| 0602 | Medicamentos estratégicos |
| 0603 | Medicamentos de âmbito hospitalar e urgência |
| 0604 | Componente Especializado da Assitencia Farmaceutica |
| 0701 | Órteses, próteses e materiais especiais não relacionados ao ato cirúrgico |
| 0702 | Órteses, próteses e materiais especiais relacionados ao ato cirúrgico |
| 0801 | Ações relacionadas ao estabelecimento |
| 0802 | Ações relacionadas ao atendimento |
| 0803 | Autorização / Regulação |

Fonte: Elaborada pelo autor (2020)

Ressalta-se também que os dados listados na tabela acima estão disponibilizados no datacenter do TCU no esquema BDU_SECEXSAUDE_SIH.dbo.TB_SUBGR.

Ainda sobre o código SIGTAP, importante enfatizar que, embora os códigos sejam os mesmos usados tanto no SIA quanto no SIH, para fins de clusterização deste trabalho, optou-se por manter uma diferenciação entre eles, a fim de poder contabilizar isoladamente os procedimentos de cada um. Assim sendo, definiu-se a seguinte rotulação, que foi utilizada durante todo o processo de mineração:

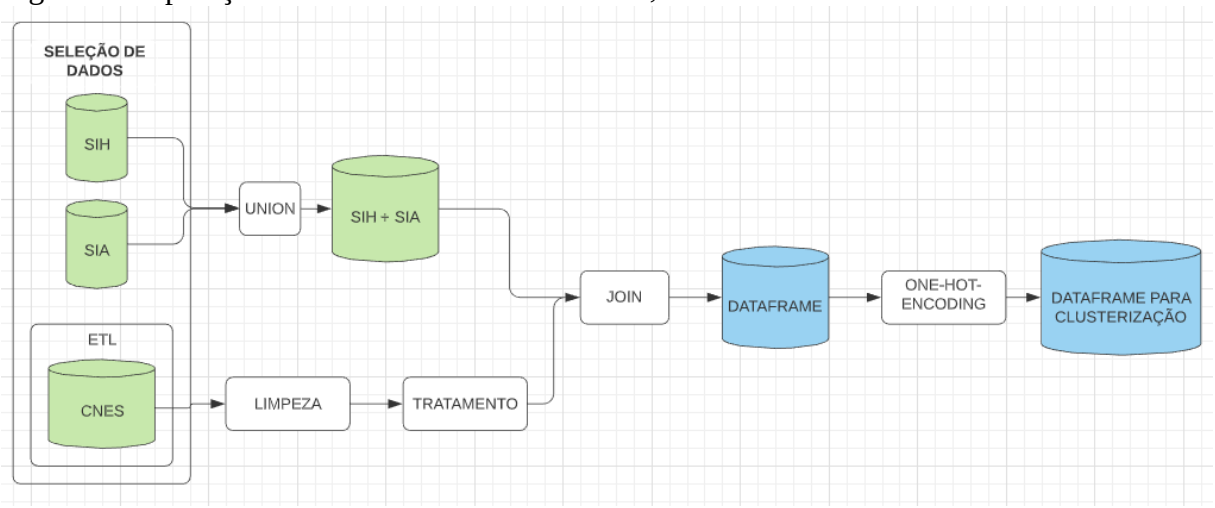
- Procedimentos do SIH - iniciando com o prefixo: '1 – '
- Procedimentos do SIA - iniciando com o prefixo: '2 – '

A título de exemplo, o procedimento "Coleta de material", cujo código de Subgrupo é 0201, realizado numa determinada unidade de saúde pelos dois tipos de atendimentos (SIH e SIA), receberão rótulos diferenciados neste trabalho, sendo: '1 – 0201' para o SIH, '2 – 0201' para o SIA. Em entrevista com a Unidade de Negócio, vislumbrou-se essa necessidade justamente para preservar o peso que cada atendimento possui isoladamente na proporção total da unidade de saúde.

## 6.3 INTEGRAÇÃO DOS DADOS

Importante lembrar que todas as tabelas utilizadas nesse trabalho, obtidas nos sistemas CNES, SIH e SIA, possuem o campo identificador da unidade de saúde (CNES) em seus dados, possibilitando a integração entre elas. Assim sendo, foram realizadas diversas operações nas tabelas de forma a resultar num conjunto de dados único e admissível para mineração. A Figura 9 – Operações realizadas nas tabelas CNES, SIH e SIA, a seguir, mostra, de forma visual, a sequência de ações realizadas durante a integração dos dados.

Figura 9 – Operações realizadas nas tabelas CNES, SIH e SIA




Fonte: Elaborada pelo autor (2020)

Esclarecendo os passos ilustrados na Figura 9 – Operações realizadas nas tabelas CNES, SIH e SIA, temos:

1) Primeiro, foram realizadas as seguintes seleções nos dados, diretamente do Labcontas:
    a. dados do SIH e do SIA individualmente, para o período de 2018, agrupados por:
        i. Campo CNES
        ii. Quatro primeiros dígitos do código SIGTAP (ou seja, SUBGRUPO).
            - SIH → campo PROC_REA
            - SIA → campo PA_PROC_ID
    b. dados do CNES, no mês de dezembro de 2018, de unidades que possuam vínculo com o SUS, cujo tipo de estabelecimentos é igual a:
        i. Hospital geral (TP_UNID = 05)
        ii. Hospital Especializado (TP_UNID = 07)
        iii. Hospital Dia (TP_UNID = 62)
2) Na sequência, foi realizado UNION entre os dados obtidos no SIH e SIA, de forma a construir um *dataframe* contendo a totalidade dos procedimentos realizados;
3) Depois, foi realizado a LIMPEZA dos dados obtidos na seleção do CNES;
4) Em seguida, foi realizado o TRATAMENTO dos dados resultantes do processo de LIMPEZA do CNES;
5) Na sequência, foi realizado JOIN entre os dados resultantes do UNION entre SIH e SIA e os dados resultantes do TRATAMENTO do CNES, resultando num *dataframe* único contendo todos os dados necessários.

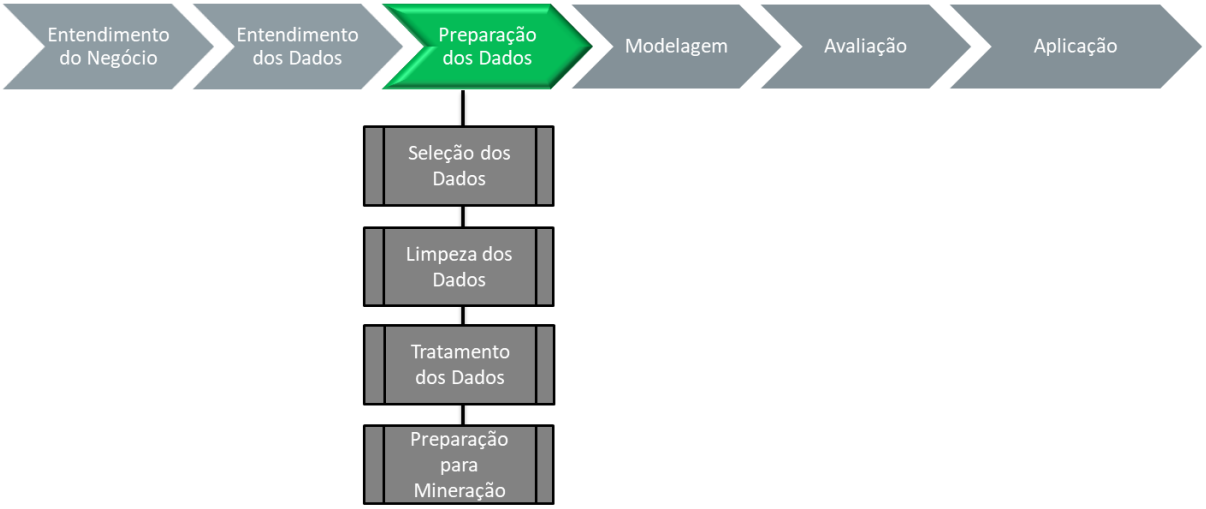
6) Por fim, foi realizado operação de *One-Hot-Encoding* para geração do *dataframe* final para clusterização.

Os detalhes das seleções dos dados estão pormenorizados no item PREPARAÇÃO DOS DADOS, a seguir.

# 7 PREPARAÇÃO DOS DADOS

A etapa de preparação de dados é usualmente a mais demorada e uma das mais importantes do *CRISP-DM* (IBM, 2014). De fato, o tempo gasto nesta etapa superou os 60% do total gasto no projeto. Esta etapa envolveu várias atividades, mostradas na Figura 10 - Detalhamento da fase “Preparação dos Dados” do *CRISP-DM*, abaixo.

Figura 10 - Detalhamento da fase “Preparação dos Dados” do *CRISP-DM*




Fonte: Elaborada pelo autor (2020)

## 7.1 SELEÇÃO DOS DADOS

Conforme já mencionado, seguindo orientação da Unidade Técnica, optou-se por limitar o escopo do trabalho apenas às unidades de saúde com vínculo ao SUS, cujo tipo de estabelecimento fosse dos tipos: Hospital Geral, Hospital Especializado ou Hospital Dia.

O período selecionado foi:

- dezembro de 2018, para o banco de dados do CNES;
- ano inteiro de 2018 para SIH e SIA.

### 7.1.1 Seleção de dados do Sistema CNES

Para seleção dos dados, uma vez que as metas de internalização de dados estavam concluídas, foi realizado acesso diretamente no Labcontas (esquema BDU_SECEXSAUDE_CNES.dbo). Das treze tabelas disponíveis, apenas quatro foram utilizadas. O total de registros selecionados, usando os critérios e filtros definidos na Figura 9 – Operações realizadas nas tabelas CNES, SIH e SIA, totalizaram **4.847 unidades**, agrupadas da seguinte forma:

- HOSPITAL GERAIS 4.075 registros
- HOSPITAL ESPECIALIZADO 576 registros
- HOSPITAL DIA 196 registros

Os campos selecionados do esquema BDU_SECEXSAUDE_CNES.dbo foram os seguintes:

Tabela 7 - Campos utilizados das tabelas internalizadas no CNES

| CAMPO | TABELA | TIPO | DESCRIÇÃO |
|---|---|---|---|
| CNES | ST | CHAR (7) | Número nacional do estabelecimento de saúde |
| TP_UNID | ST | CHAR (2) | Tipo de unidade (estabelecimento). Selecionados os tipos: 05, 07 e 62 |
| VINC_SUS | ST | CHAR (1) | Vínculo com SUS: 1-Sim 0-Não |
| QT_SUS | LT | NUMERIC (4) | Quantidade de leitos destinados ao SUS pela unidade de saúde |
| CODEQUIP | EQ | CHAR (2) | Código do equipamento |
| QT_EXIST | EQ | CHAR (4) | Quantidade existente do equipamento |
| CBO | PF | CHAR (6) | Código da especialidade do profissional |

Fonte: Elaborada pelo autor (2020)

### 7.1.2 Seleção de dados do SIH

Os dados selecionados do sistema SIH, obtidos no esquema BDU_SECEXSAUDE_SIH.dbo, utilizando os critérios e filtros definidos na Figura 9 – Operações realizadas nas tabelas CNES, SIH e SIA, **totalizaram 45.609 registros**. Os campos estão listados na tabela abaixo.

Tabela 8 - Campos utilizados das tabelas internalizadas do SIH

| CAMPO | TABELA | TIPO | DESCRIÇÃO |
|---|---|---|---|
| CNES | RD | CHAR (7) | Número nacional do estabelecimento de saúde |
| PROC_REA | RD | CHAR (10) | Código do procedimento (SIGTAP) |
| N_AIH | RD | CHAR (10) | Código da internação |
| VAL_TOT | RD | NUMERIC (13,2) | Valor total gasto com o procedimento |

Fonte: Elaborada pelo autor (2020)

### 7.1.3 Seleção de dados do SIA

Da mesma forma que o SIH, os dados do sistema SIA selecionados do Labcontas no esquema BDU_SECEXSAUDE_SIA.dbo, utilizando os critérios e filtros definidos na Figura 9

– Operações realizadas nas tabelas CNES, SIH e SIA, totalizaram **277.086 registros**. Os campos estão listados na tabela abaixo.

Tabela 9 - Campos utilizados das tabelas internalizadas do SIA

| CAMPO | TABELA | TIPO | DESCRIÇÃO |
|---|---|---|---|
| CNES | SIA_PA | CHAR (7) | Número nacional do estabelecimento de saúde |
| PA_PROC_ID | SIA_PA | CHAR (10) | Código do procedimento realizado (SIGTAP) |
| PA_QTDAPR | SIA_PA | INT | Quantidade do procedimento realizado |
| PA_VALAPR | SIA_PA | FLOAT | Valor total gasto com o procedimento realizado |

Fonte: Elaborada pelo autor (2020)

### 7.1.4 Seleção de dados auxiliares

Ainda no intuito de facilitar a análise dos resultados, optou-se por selecionar alguns campos de tabelas auxiliares. A lista completa está na tabela abaixo.

Tabela 10 - Campos utilizados nas tabelas auxiliares - arquivos CNV

| CAMPO | TABELA | DESCRIÇÃO |
|---|---|---|
| NOME_FANTASIA | BDU_SECEXSAUDE_CNES.dbo.TAB_AUX_ESTABELECIMENTOS | Nome da unidade de saúde cadastrada. |
| DS_ESTAB | BDU_SECEXSAUDE_CNES.dbo.TAB_AUX_TP_ESTABELECIMENTOS | Tipo do estabelecimento de saúde |
| CO_SUB_GRU, | BDU_SECEXSAUDE_SIH.dbo.TB_SUBGR | Código do SUBGRUPO do procedimento SIGTAP |
| NO_SUB_GRU | BDU_SECEXSAUDE_SIH.dbo.TB_SUBGR | Nome do SUBGRUPO do procedimento SIGTAP |

Fonte: Elaborada pelo autor (2020)

## 7.2 LIMPEZA DOS DADOS

Como parte de limpeza de dados, foram eliminados, dos dados selecionados das tabelas do CNES, os valores inconsistentes e/ou nulos que poderiam causar ruído à análise de mineração. Os seguintes critérios foram adotados, com base em entrevistas com Unidade Técnica.

- Unidades de Saúde com quantidade de LEITOS_SUS nulo ou zero: Considerou-se que unidades de saúde do tipo “hospital” com leitos “nulos” ou com valor “0” são inconsistentes, e que a presença desses valores na base ocorreu, provavelmente, por erro de preenchimento. Com isso, optou-se por eliminá-los, para a clusterização deste trabalho. O número de registros encontrados com esse critério foi de 129.

- Unidades de saúde com quantidade de médicos “nulo” ou “zero”: Foram excluídas, pelos mesmos motivos expostos acima, unidades de saúde sem nenhum profissional de medicina ou de enfermagem alocados. O número de registros com esse tipo de dado foi de 24.
- Unidades de saúde que realizaram menos de 365 procedimentos ao ano, ou seja, 1 atendimento por dia, em média. Optou-se por esse filtro para eliminar da base unidades de saúde com porte insignificante. Adicionalmente, este filtro também pôde automaticamente excluir unidades de saúde cujos dados foram registrados de forma incompleta. O número de unidades excluídas por intermédio deste filtro foram de 324.

O total de registros obtidos da base de dados do CNES, após limpeza dos dados, foi de **4370 unidades**.

## 7.3 TRATAMENTO DOS DADOS

Após coleta inicial dos dados, foi necessária a criação de diversos campos para auxiliar no entendimento dos dados. Estes campos foram utilizados de alguma forma no decorrer do projeto nas suas diversas iterações, mencionadas no item 8 - MODELAGEM deste documento. A lista completa de campos derivados criados no decorrer do trabalho está na Tabela 11 - Campos criados para análise de mineração, abaixo.

Tabela 11 - Campos criados para análise de mineração

| CAMPO CRIADO | TABELA ORIGEM | CAMPO ORIGEM | SIGNIFICADO DO CAMPO CRIADO |
|---|---|---|---|
| QT_LEITOS_TOTAIS | LT | QT_EXIST | Somatório da quantidade de leitos disponíveis na unidade de saúde. Obtido pela fórmula sum(QT_EXIST) |
| QT_LEITOS_SUS | LT | QT_SUM | Somatório da quantidade de leitos destinados ao SUS de uma unidade de saúde. Obtido pela fórmula sum(QT_SUM) |
| QT_CONSULTORIOS | ST | QTINST01 a QTINST37 | Somatório das salas ou consultórios de uma unidade de saúde. Obtido pelo somatório dos campos QTINST01 a QTINST37 |
| QTD_EQ_TOTAIS | EQ | QT_EXIST | Somatório da quantidade de equipamentos de uma unidade de saúde. Obtido pela fórmula sum(QT_EXIST) |
| QTD_EQ_DIAG_ IMAGEM | EQ | QT_EXIST | Somatório da quantidade de equipamentos de imagem de uma unidade de saúde. . Obtido pela fórmula sum(QT_EXIST) where TIPEQUIP=1 |
| QTD_PROFISSIONAIS | PF | CBO | Somatório da quantidade de profissionais registrados em uma unidade de saúde. Obtido pela fórmula sum(CBO) |

| CAMPO CRIADO | TABELA ORIGEM | CAMPO ORIGEM | SIGNIFICADO DO CAMPO CRIADO |
|---|---|---|---|
| QTD_HORAS_MEDICAS | PF, TAB_AUX_CBO | CBO, DS_CBO | Somatório de horas MÉDICAS destinadas ao SUS registrados em uma unidade de saúde |
| QTD_HORAS_ENFERMAGEM | PF, TAB_AUX_CBO | CBO, DS_CBO | Quantidade de horas de ENFERMAGEM destinadas ao SUS registrados em uma unidade de saúde |
| QTD_PROC | SIA e SIH separadamente | PROC_REA (SIA), PA_PROC_ID e PA_QTDAPR (SIH) | Quantidade de procedimentos realizados por SUBGRUPO em uma unidade de saúde |
| VALOR_PROC | SIA e SIH separadamente | VAL_TOT (SIA), PA_VALAPR (SIH) | Valor gasto nos procedimentos por SUBGRUPO em uma unidade de saúde |
| QTD_TOTAL_PROC_POR_ATEND | SIA e SIH separadamente | PROC_REA (SIA), PA_PROC_ID e PA_QTDAPR (SIH) | Quantidade total de procedimentos realizados por tipo de atendimento (SIA e SIH separadamente) por unidade de saúde |
| VALOR_TOTAL_POR_ATEND | SIA e SIH separadamente | VAL_TOT (SIA), PA_VALAPR (SIH) | Valor total gasto nos procedimentos por tipo de atendimento (SIA e SIH separadamente) e por unidade de saúde |
| QTD_TOTAL_PROCEDIMENTOS | SIA e SIH somados | PROC_REA (SIA), PA_PROC_ID e PA_QTDAPR (SIH) | Quantidade total de procedimentos realizados por unidade de saúde |
| VALOR_TOTAL_PROCEDIMENTOS | SIA e SIH somados | VAL_TOT (SIA), PA_VALAPR (SIH) | Valor total gasto nos procedimentos por unidade de saúde |

Fonte: Elaborada pelo autor (2020)

## 7.4 PREPARAÇÃO PARA MINERAÇÃO

Ainda na parte de preparação dos dados, foram realizadas diversas atividades no intuito de preparar o *dataframe* final para a execução da modelagem. As atividades estão detalhadas nos itens a seguir.

### 7.4.1 Criação de campo calculado usado na mineração

Foi criado um campo calculado para ser usado para clusterização. Este campo, denominado **PERCENTUAL_VALOR,** foi obtido pelo cálculo das proporções entre o valor do procedimento realizado individualmente e o total gasto na unidade de saúde. A fórmula aplicada foi:

```
PERCENTUAL_VALOR  = VALOR_PROC / VALOR_TOTAL_PROCEDIMENTOS
```

Equação 1 – Campo usado na Mineração

Um extrato do *dataframe*, após coleta, limpeza, tratamento e UNION dos dados, está listado na figura abaixo.

Tabela 12 – Extrato do *dataframe* após tratamento dos dados

| CAMPOS OBTIDOS DIRETO DA BASE | | | | CAMPOS CRIADOS | |
|---|---|---|---|---|---|
| **CNES** | **CO_SUB_GRU** | **QTD_PROC** | **VALOR_PROC** | **VALOR_TOTAL_PROCEDIMENTOS** | **PERCENTUAL_VALOR** |
| 0000035 | 1 - 0301 | 14 | 847,08 | 1391025,73 | 0,000609 |
| 0000035 | 1 - 0303 | 225 | 108534,41 | 1391025,73 | 0,078025 |
| 0000035 | 1 - 0304 | 12 | 2160,67 | 1391025,73 | 0,001553 |
| 0000035 | 1 - 0305 | 11 | 2929,5 | 1391025,73 | 0,002106 |
| 0000035 | 1 - 0308 | 2 | 329,9 | 1391025,73 | 0,000237 |
| 0000035 | 1 - 0401 | 2 | 332,22 | 1391025,73 | 0,000239 |
| 0000035 | 1 - 0407 | 125 | 71329,96 | 1391025,73 | 0,051279 |
| 0000035 | 1 - 0408 | 4 | 2469,58 | 1391025,73 | 0,001775 |
| 0000035 | 1 - 0409 | 56 | 19225,66 | 1391025,73 | 0,013821 |
| 0000035 | 2 - 0201 | 442 | 7017 | 1391025,73 | 0,005044 |
| 0000035 | 2 - 0204 | 4532 | 36373,46 | 1391025,73 | 0,026149 |
| 0000035 | 2 - 0209 | 190 | 23546,7 | 1391025,73 | 0,016928 |
| 0000035 | 2 - 0211 | 1078 | 6961,61 | 1391025,73 | 0,005005 |
| 0000035 | 2 - 0301 | 169551 | 1000421,68 | 1391025,73 | 0,719197 |
| 0000035 | 2 - 0401 | 8872 | 94939,26 | 1391025,73 | 0,068251 |
| 0000035 | 2 - 0407 | 456 | 13607,04 | 1391025,73 | 0,009782 |

Fonte: Elaborada pelo autor (2020)

Na Tabela 12 – Extrato do *dataframe* após tratamento dos dados, o campo CO_SUB_GRU (segunda coluna da tabela) contém os códigos dos subgrupos de procedimentos SIGTAP realizados. Ressalta-se que os que iniciam com ‘1 –‘ são os procedimentos obtidos do SIH, que se referem às internações hospitalares. Os que iniciam com ‘2 – ‘ são os procedimentos obtidos do SIA, que se referem aos atendimentos ambulatoriais.

O campo VALOR_PROC (quarta coluna) contém os valores gastos em cada procedimento. O total gasto na unidade como um todo está listado no campo VALOR_TOTAL_PROCEDIMENTOS (quinta coluna). Analisando a tabela acima, constata-se que a unidade de saúde de número 0000035 gastou, em valores monetários, o correspondente a R$ 1.391.025,73 no ano de 2018, entre procedimentos SIH e SIA.

O campo PERCENTUAL_VALOR (última coluna) foi criado para ser usado na clusterização, e contém a proporção do valor gasto na unidade por procedimento realizado, conforme mostrado na Equação 1 – Campo usado na Mineração. Importante notar que o somatório dos valores do campo PERCENTUAL_VALOR de cada unidade retorna sempre 1.

### 7.4.2 Técnica de One-Hot-Encoding

Ainda na fase de Preparação de Dados, foi necessária a aplicação da técnica de *One-Hot-Encoding* no *dataframe* antes da aplicação dos algoritmos de modelagem, tendo em vista que os modelos escolhidos neste trabalho utilizam campos numéricos apenas. Esta técnica consiste em transformar as variáveis categóricas em colunas, eliminando as demais não utilizadas na mineração. A variável categórica transformada em coluna foi a CO_SUB_GRU (segunda coluna da Tabela 12), e os campos numéricos foram descartados, com exceção do PERCENTUAL_VALOR, usado na clusterização. A Tabela 13 - Exemplo do *dataframe* final usado no modelo de clusterização, mostra um extrato do *dataframe* após a aplicação desta técnica.

Tabela 13 - Exemplo do *dataframe* final usado no modelo de clusterização

| **CO_SUB_GRU** | **1 - 0201** | **1 - 0209** | **1 - 0211** | **1 - 0301** | **1 - 0303** | **...** | **2 - 0210** | **2 - 0211** | **2 - 0212** |
|---|---|---|---|---|---|---|---|---|---|
| CNES | | | | | | ... | | | |
| 0000035 | 0.000000 | 0.000000 | 0.000000 | 0.000609 | 0.078025 | ... | 0.000000 | 0.003741 | 0.000000 |
| 0000094 | 0.000000 | 0.000000 | 0.000000 | 0.000399 | 0.040798 | ... | 0.000000 | 0.024999 | 0.000000 |
| 0000396 | 0.000487 | 0.000105 | 0.000000 | 0.003099 | 0.145922 | ... | 0.000197 | 0.007382 | 0.000455 |
| 0000418 | 0.000000 | 0.000000 | 0.000000 | 0.000050 | 0.394458 | ... | 0.000000 | 0.016943 | 0.000000 |
| 0000426 | 0.000088 | 0.000000 | 0.000962 | 0.002137 | 0.379934 | ... | 0.000041 | 0.035544 | 0.000244 |

Fonte: Elaborada pelo autor (2020)

Nota-se que, após aplicação do *One-Hot-Encoding*, cada linha do *dataframe* mostrado na Tabela 13 - Exemplo do *dataframe* final usado no modelo de clusterização, passou a conter uma única unidade de saúde (CNES), sendo que os procedimentos realizados foram transpostos para as colunas. Cada coluna representa um tipo de procedimento realizado (CO_SUB_GRU). Na tabela em questão, é mostrado apenas uma parte das colunas, de um total de 83. Os valores de cada célula correspondem ao campo PERCENTUAL_VALOR, e assumem valores de zero quando não possuírem procedimento algum realizado na unidade.

Neste ponto do trabalho, o *dataframe* resultou em **4370 linhas e 83 colunas**, estando apto para ser clusterizado na modelagem.

# 8 MODELAGEM

Nesta etapa do *CRISP-DM*, os dados preparados na etapa anterior são executados pelos algoritmos de mineração. É aqui que os resultados começam a esclarecer os problemas de negócios apresentados durante o entendimento dos negócios. A modelagem normalmente é realizada em diversas iterações. É raro, em um projeto de dados, que a pergunta da mineração seja respondida satisfatoriamente com um único modelo e uma única execução (IBM, 2014).

De fato, diversas iterações ocorreram no decorrer deste projeto até que se chegasse à versão final. A título de registro, a Tabela 14 - Iterações realizadas durante a modelagem, abaixo, mostra as quatro principais iterações realizadas durante o trabalho, cujos resultados nortearam a conclusão do presente documento.

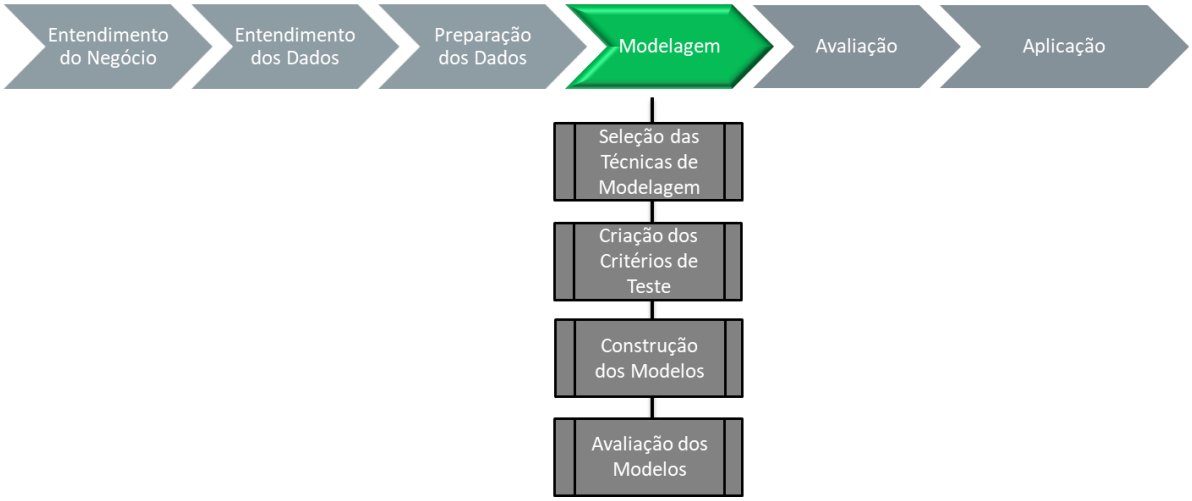
Tabela 14 - Iterações realizadas durante a modelagem

| ITERAÇÃO | DADOS UTILIZADOS PARA MINERAÇÃO | RESULTADOS OBTIDOS | AVALIAÇÃO |
|---|---|---|---|
| 1ª | QT_LEITOS_SUS<br>QT_CONSULTORIOS<br>QTD_EQ_DIAG_IMAGEM<br>QTD_PROFISSIONAIS<br>QTD_HORAS_MEDICAS<br>QTD_HORAS_<br>ENFERMAGEM | Separação em clusters por porte da unidade de saúde. | Constatou-se que esta não seria a melhor estratégia pois o porte da unidade não necessariamente traduz o seu perfil de atendimento. |
| 2ª | QTD_TOTAL_PROC_<br>POR_ATEND / QTD_PROC<br>(SIA e SIH somados) | Separação em clusters por perfil de SUBGRUPO de procedimento considerando os QUANTITATIVOS realizados no SIA e SIH somados. | Constatou-se que, embora os resultados fossem melhores que na 1ª iteração, ainda não estavam satisfatórios, tendo em vista que o SIA possui quantidade muito superior de procedimentos realizados que o SIH, refletindo em um peso maior ao SIA na ponderação dos perfis dos clusters. |
| 3ª | QTD_TOTAL_PROC_<br>POR_ATEND / QTD_PROC<br>(SIA e SIH separados) | Separação em clusters por perfil de SUBGRUPO de procedimento considerando os QUANTITATIVOS realizados no SIA e SIH separadamente. | Constatou-se que ainda não seria a melhor estratégia, pois o uso da variável QUANTITATIVO gerava um viés alto na separação dos clusters. |
| 4ª | PERCENTUAL_ VALOR<br>(SIA e SIH separados) | Separação em clusters por perfil de SUBGRUPO de procedimento, considerando os VALORES gastos pelo SIA e SIH separadamente. | Constatou-se a melhor opção de clusterização, tendo em vista que os clusters separavam por perfil de atendimento, considerando os PERCENTUAL DOS VALORES gastos, o que representa maior realidade. Esta foi a opção escolhida e relatada neste trabalho. |

Fonte: Elaborada pelo autor (2020)

A Figura 11 - Detalhamento da fase “Modelagem” do *CRISP-DM*, abaixo, mostra as atividades realizadas nessa etapa do *CRISP-DM*.

Figura 11 - Detalhamento da fase “Modelagem” do *CRISP-DM*




Fonte: Elaborada pelo autor (2020)

## 8.1 SELEÇÃO DA TÉCNICA DE MODELAGEM

Primeiramente, cabe salientar que a técnica de mineração empregada neste trabalho – Clusterização, se encaixa na categoria da técnica “não supervisionada”, o que indica que os dados não possuem rótulo, isto é, não se sabe de antemão quais unidades de saúde pertencem a quais grupos. O algoritmo é o responsável pela separação das unidades, utilizando como base os padrões contidos nos próprios dados.

Como a tarefa de agrupamento é subjetiva, os meios usados para atingir esse objetivo são muitos. Cada algoritmo de clusterização segue um conjunto diferente de regras para definir a ‘similaridade’ entre os pontos de dados. De fato, existem mais de uma centena de algoritmos de clusterização conhecidos, mas poucos deles são usados popularmente (DATASCIENCE, 2020).

Segundo MAXWEL (2014), as categorias principais dos modelos de clusterização são: Métodos Hierárquicos; Métodos Particionais; Métodos baseados em densidade; Métodos baseados em grade; Métodos baseados em modelos; Métodos baseados em Redes Neurais; Métodos baseados em Lógica Fuzzy; Métodos baseados em Kernel; Métodos baseados em Grafos e Métodos baseados em Computação Evolucionária.

Como destacado por HAN J (2001), alguns algoritmos de Clusterização integram as ideias de vários outros, então, algumas vezes, é difícil classificar um dado algoritmo como unicamente pertencendo a somente uma categoria de método de Clusterização.

Neste trabalho, tendo em vista a escalabilidade dos dados bem como o tempo necessário de processamento, optou-se por testar quatro tipos de algoritmos, a seguir listados:

- Métodos Particionais - K-means
- Métodos baseados em densidade - Mean Shift
- Métodos baseados em densidade - DBScan
- Métodos Hierárquicos - AgglomerativeClustering

## 8.2 CRIAÇÃO DOS CRITÉRIOS DE TESTE

Antes da construção dos modelos propriamente dito, foi necessário definir os critérios de teste que seriam utilizados para avaliar o resultado dos modelos. Em entrevista com a Unidade Técnica, identificou-se dois critérios que nortearam a escolha do melhor modelo a ser usado. Os critérios foram:

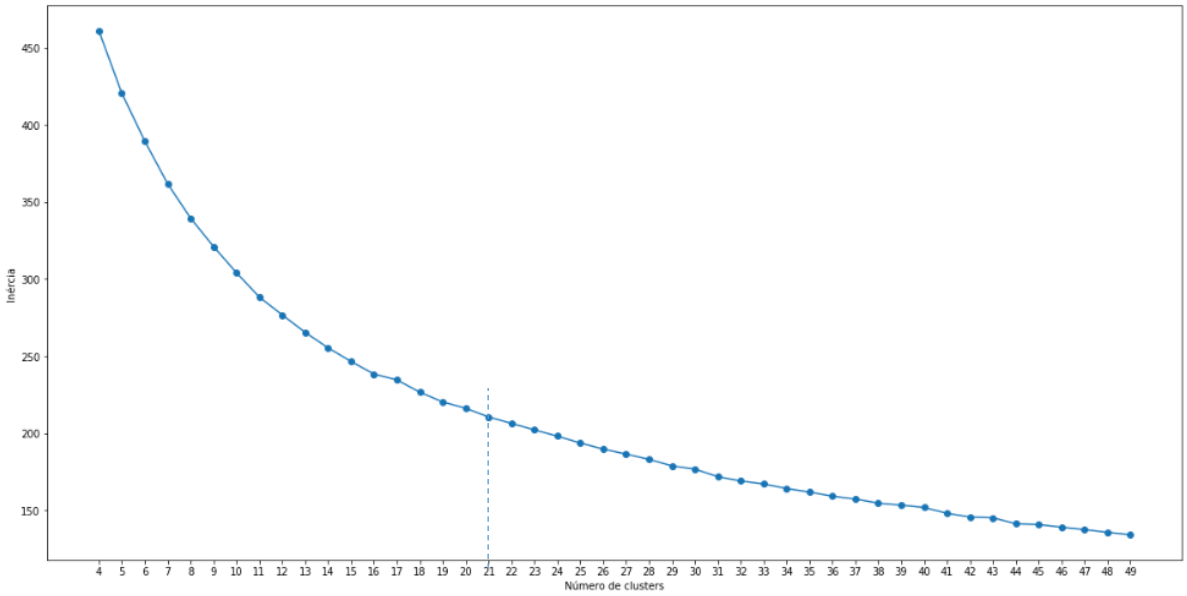
- número mínimo de clusters – ficou definido que a quantidade mínima aceitável de clusters seria de **quatro**, tendo em vista que trabalhos anteriores realizados, a exemplo do Banco Mundial (BM, 2019), utilizaram esta mesma quantidade de grupos.
- quantidade de membros em cada cluster – ficou definido que, em que pese a utilização da técnica DEA, onde é desejável que o número de membros de cada cluster não seja inferior ao número de variáveis (input + output) multiplicada por três, o número de membros de cada cluster não poderia ser inferior a **trinta**, já que o número de variáveis previstas pela SecexSaúde na utilização do DEA seria de dez (sete inputs e três outputs) (WILLIAM W. COOPER, 2011).

Além dos critérios de testes definidos, foram utilizados três recursos adicionais para identificar a quantidade ideal de clusters: *Elbow Method, Silhouette coeficiente* e análise visual dos dados.

### *8.2.1 ElbowMethod*

A *Elbow Method* ajuda a selecionar o número ideal de clusters ajustando o modelo com uma faixa de valores para K. Se o gráfico de linhas se assemelhar a um braço, o "cotovelo" (o ponto de inflexão na curva) indicaria o ponto ideal do modelo subjacente (DEVELOPERS, 2016). A figura abaixo mostra o resultado do *Elbow Method* aplicada para a técnica K-means no *dataframe* usado para este trabalho.

Figura 12 - *Elbow Method* aplicado no *dataframe*




Fonte: Elaborada pelo autor (2020)

Analisando a Figura 12 - *Elbow Method* aplicado no *dataframe*, observa-se uma inflexão bem suave na curva a cada ponto do gráfico. Entretanto, percebe-se desvios ainda menores após o cluster 21, indicando uma possível estabilidade no número de clusters a partir daí. Embora a figura 12 tenha indicado que, a priori, o número ideal de clusters estaria entre os pontos 4 e 21, optou-se por descartar esta técnica e partir para a análise do coeficiente de *Silhouette*,

### 8.2.2 Coeficiente de *Silhouette*

Não havendo um ponto claro na Figura 12 - *Elbow Method* aplicado no *dataframe* que indicasse o número ideal de clusters, optou-se por realizar outra análise, desta vez usando o *Silhouette coefficient*. Esta técnica ajuda a identificar quão semelhante um objeto é ao seu próprio cluster (coesão) em comparação com outros clusters (separação). O valor do *silhouette* varia de -1 a +1, onde um valor alto indica que o objeto é bem correspondido ao seu próprio cluster e mal correspondido aos clusters vizinhos.

Esta técnica foi aplicada para todos os modelos executados, e seus resultados serão mostrados mais adiante neste documento.

O cálculo do *Silhouette coeficiente* consiste em definir a distância média de um ponto para todos os outros pontos em seu cluster *a(i),* como também a distância média até os pontos do cluster mais próximo *b(i)* (MEDIUM, 2020). O *Silhouette coefficient s(i)* é calculado da seguinte forma:

$$s(i) = (b(i) - a(i)) / max(b(i), a(i))$$

Equação 2 – *Silhouette Coefficient*

### 8.2.3 Análise Visual dos Dados

Visando facilitar a visualização dos dados, e com isso permitir uma ferramenta adicional de análise dos clusters pós a modelagem, foi realizada a redução da dimensionalidade dos dados do *dataframe*. A técnica escolhida foi o PCA - *Principal Component Analysis*.

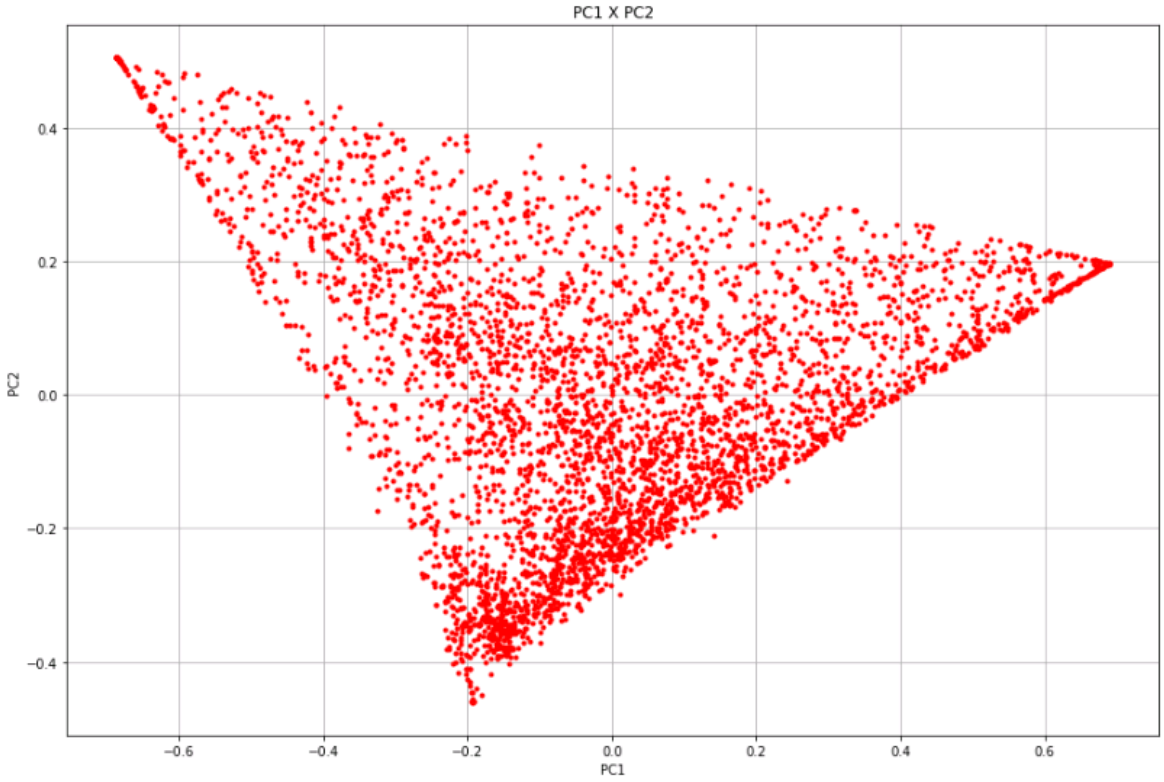
A técnica PCA consiste na análise estatística multivariada que transforma linearmente um conjunto original de variáveis, inicialmente correlacionadas entre si, num conjunto substancialmente menor não correlacionadas, mantendo a maior parte da essência da informação do conjunto original (KUANG HONGYU, 2015).

O objetivo principal desta técnica é reduzir a dimensionalidade da amostra, possibilitando identificar o quanto cada componente contribui para "explicar" o dado como um todo. O PCA foi aplicado nesta análise para reduzir as dimensões do conjunto de dados para duas colunas apenas (pc1 e pc2), possibilitando a visualização dos dados em gráfico e, com isso, permitir a comparação dos resultados nos diferentes clusters.

Importante salientar que a redução da dimensionalidade descrita nesta etapa se deu exclusivamente para fins de visualização dos dados. Os algoritmos de modelagem em si utilizaram o *dataframe* original, com a totalidade das colunas.

Após aplicação do PCA, foi possível a exibição dos dados. A Figura 13 – Plotando *dataframe* antes da clusterização, a seguir, mostra uma fotografia dos dados, após aplicação do PCA e antes da execução da modelagem.

Figura 13 – Plotando *dataframe* antes da clusterização

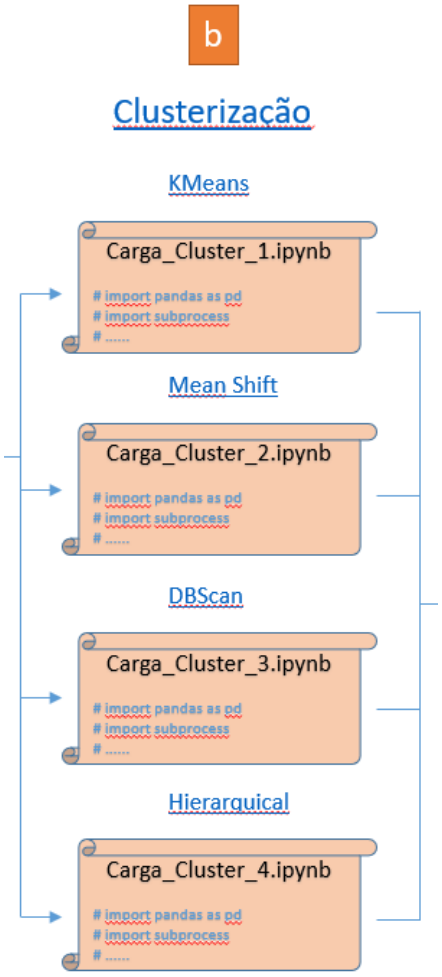



Fonte: Elaborada pelo autor (2020)

## 8.3 CONSTRUÇÃO DO MODELO

Esta etapa consiste na construção do modelo propriamente dito. Esta atividade contemplou a segunda meta de mineração ilustrada na Figura 4 – Produtos desenvolvidos no trabalho, mostrada a seguir.

Figura 14 – Clusterização




Fonte: Elaborada pelo autor (2020)

Nesta etapa, foram executados quatro diferentes algoritmos de mineração de dados, utilizando a linguagem *python* (PYTHON, 2001) e bibliotecas específicas. Ressalta-se que os modelos escolhidos se encaixam na modalidade de "não supervisionada" de clusterização, e que os dados de entrada (*dataframe*) utilizados nos quatro modelos foi exatamente o mesmo. Abaixo, é detalhado cada modelo, com os respectivos resultados obtidos.

### 8.3.1 Modelo K-means

É o mais popular e mais simples modelo de clusterização. Tem como pré-requisito que se informe, a priori, o número de clusters desejado. A partir daí o algoritmo seleciona, de forma aleatória, grupo de centroides como pontos de partida para cada cluster e, em seguida, executa cálculos iterativos (repetitivos) para otimizar as posições dos centroides. A execução interrompe quando os centroides se estabilizaram, ou seja, quando não há alteração nos centroides, ou o número definido de iterações foi alcançado (SCIENCE, 2016).

O *Silhouette coefficient* foi aplicado para o modelo K-means, variando o número de clusters de 4 a 25. Seus resultados foram ordenados de forma decrescente pelo coeficiente, e exibidos os 5 primeiros na Tabela 15, a seguir. A primeira linha corresponde ao valor mais alto de *silhouette* obtido para o k-means.

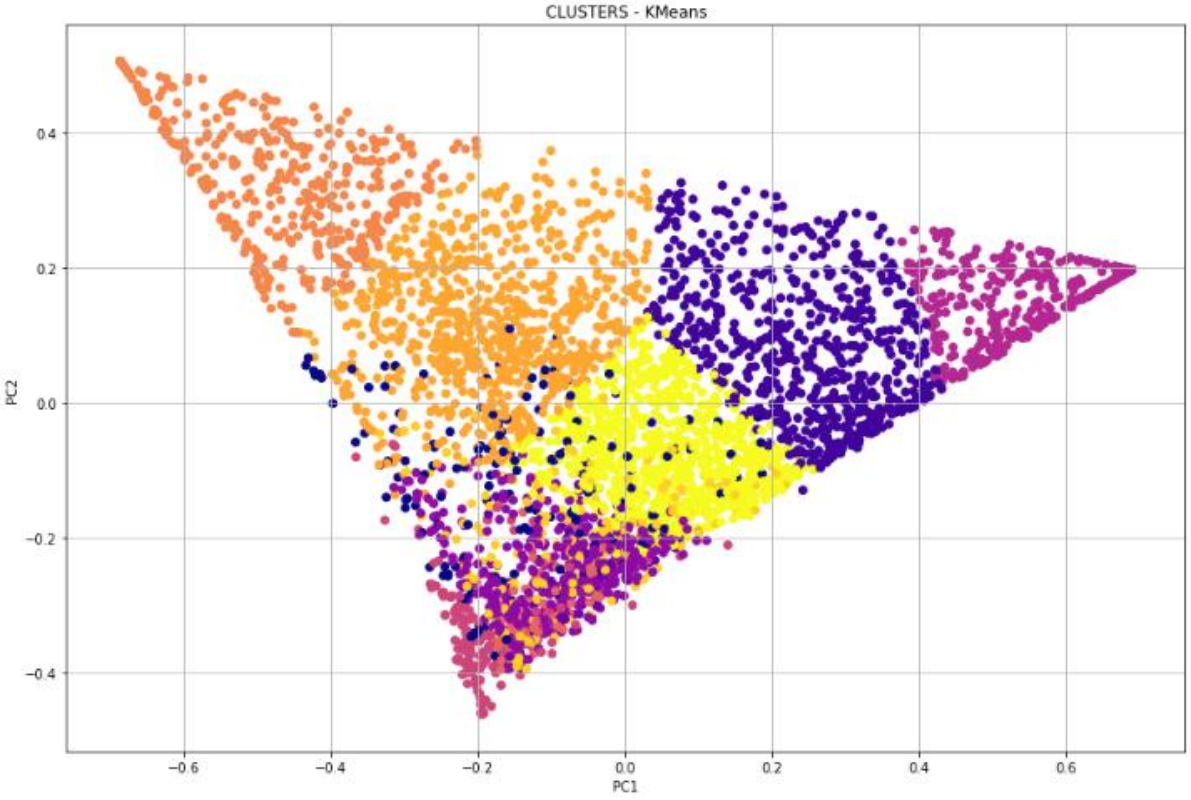
Tabela 15 - Valores de *Silhouette* para K-means

| POSIÇÃO | NÚMERO DE CLUSTERS | VALOR DE SILHOUETTE |
|---|---|---|
| 1 | 11.0 | 0.371322 |
| 2 | 9.0 | 0.359809 |
| 3 | 10.0 | 0.357319 |
| 4 | 8.0 | 0.345085 |
| 5 | 19.0 | 0.342455 |

Fonte: Elaborada pelo autor (2020)

Com base no resultado da Tabela 15 - Valores de *Silhouette* para K-means, optou-se pela escolha do número de 11 clusters para o modelo K-means. O resultado da figura abaixo mostra visualmente a separação dos clusters após aplicação do Modelo K-means.

Figura 15 – Plotando o *dataframe* após clusterização K-means




Fonte: Elaborada pelo autor (2020)

### 8.3.2 Modelo Mean Shift

Mean Shift é uma técnica de clusterização de análise de espaço para localizar os máximos de uma função de densidade, também conhecido como *seeking algorithm* (CHENG, 1995). É um algoritmo baseado em centroide, pois funciona atualizando os candidatos aos centroides com a média dos pontos em uma determinada região. Esses candidatos são então filtrados em um estágio de pós-processamento para eliminar duplicatas próximas para formar o conjunto final de centroides (SCIKIT-LEARN, 2020).

O algoritmo Mean Shift é pouco escalável, pois requer várias pesquisas de vizinhos mais próximos durante sua execução, o que pode demorar mais seu processamento. Necessita de um parâmetro denominado Bandwith, derivado de duas medidas: quantile e n_samples, as quais definem o tamanho da região a pesquisar.

Aqui, também foi aplicada a análise do *Silhouette coefficient* para variados valores de parâmetros. Os intervalos testados foram a combinação dos seguintes valores:

- Quantile: 0.1, 0.2, 0.3, 0.4 e 0.5
- N_Samples: 20, 50, 100

Os melhores resultados estão listados na tabela 16 abaixo.

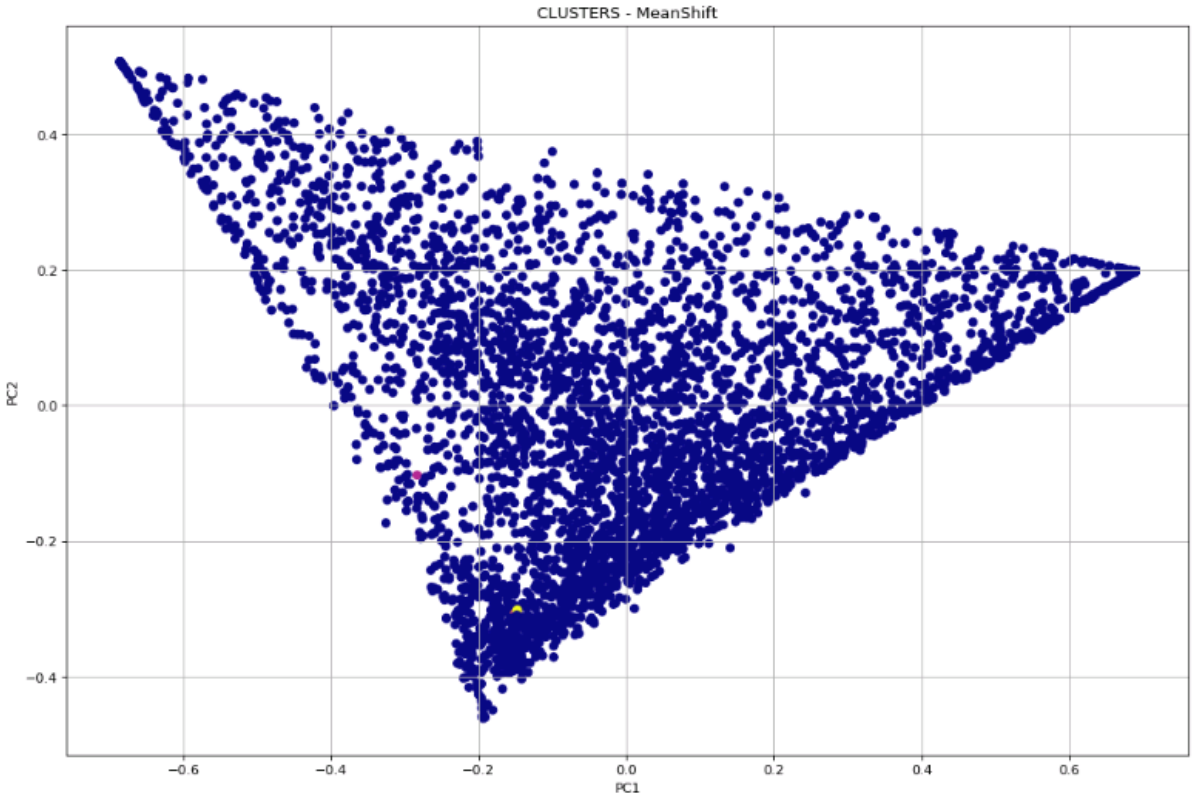
Tabela 16 - Valores de *Silhouette* para Mean Shift

| POSIÇÃO | QUANTILE | N_SAMPLE | NÚMERO DE CLUSTERS | VALOR DE SILHOUETTE |
|---|---|---|---|---|
| 1 | 0.5 | 20.0 | 6.0 | 0.535989 |
| 2 | 0.5 | 100.0 | 6.0 | 0.535989 |
| 3 | 0.4 | 100.0 | 9.0 | 0.345581 |
| 4 | 0.5 | 50.0 | 11.0 | 0.218412 |
| 5 | 0.3 | 20.0 | 16.0 | 0.212345 |

Fonte: Elaborada pelo autor (2020)

Os parâmetros escolhidos, considerando o melhor valor de *Silhouette* listado na primeira linha da Tabela 16 - Valores de *Silhouette* para Mean Shift, indicou o agrupamento em 6 clusters. A figura abaixo mostra resultado após aplicação da modelagem Mean-Shift.

Figura 16 - Plotando o *dataframe* após clusterização Mean Shift




Fonte: Elaborada pelo autor (2020)

Embora pouco perceptível, a Figura 16 - Plotando o *dataframe* após clusterização Mean Shift mostra seis clusters distintos, sendo que a grande maioria das unidades se encontram no cluster representado pela cor azul.

### 8.3.3 Modelo DBSCAN

O algoritmo DBSCAN trabalha com áreas, separando as de alta densidade das de baixa densidade. Devido a esse espectro bastante genérico, os clusters encontrados pelo DBSCAN podem ter qualquer formato. O DBSCAN utiliza o conceito de amostras principais em áreas de alta densidade. Um cluster é, portanto, um conjunto de “amostras principais”, cada uma próxima da outra e de um conjunto de amostras “não principais”. Existem dois parâmetros utilizados nesse tipo de algoritmo: min_samples e eps, que definem o que se pretende denotar “densidade” (SCIKIT-LEARN, 2019).

Da mesma forma, foi aplicada a análise do *Silhouette coefficient* para diversos valores de parâmetros. Os intervalos de parâmetros testados foram a combinação dos seguintes valores:

- EPS: 0.1, 0.2, 0.3, 0.4 e 0.5
- Min_Samples: 5, 10, 20, 50 e 100

Os melhores resultados estão listados na tabela 17, abaixo.

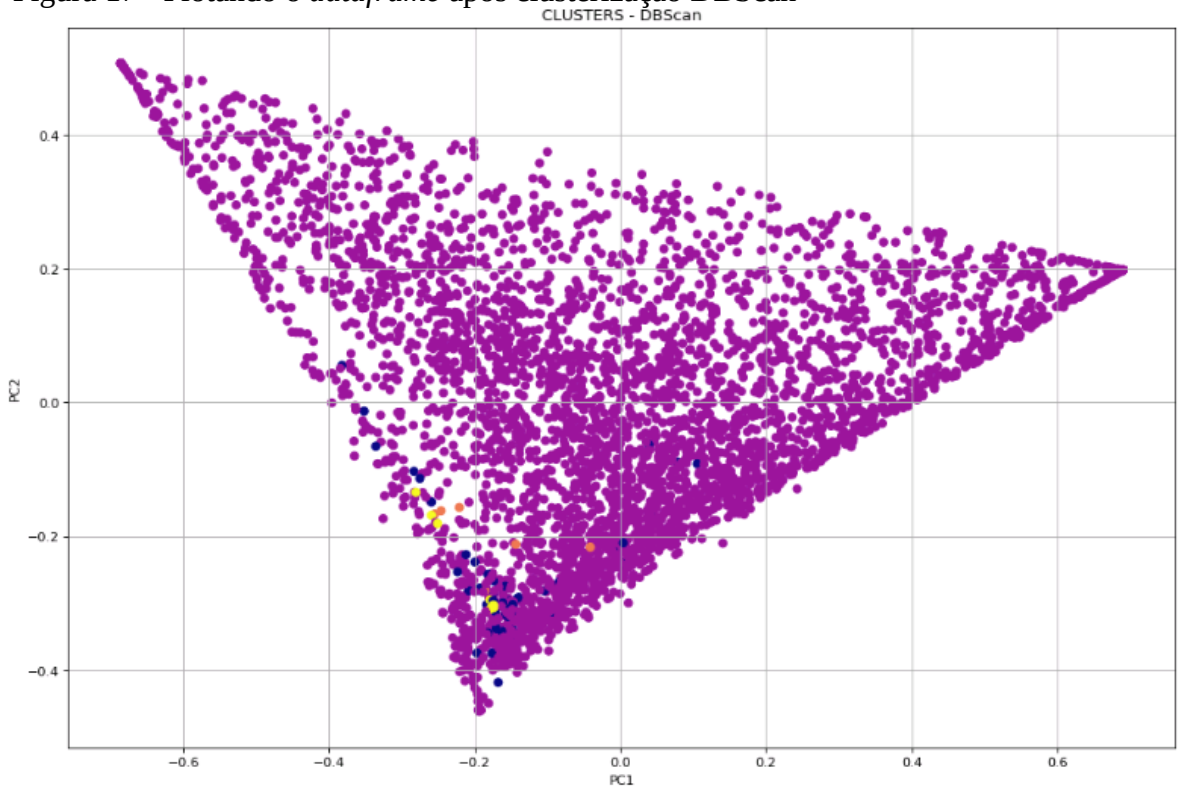
Tabela 17 - Valores de *Silhouette* para DBScan

| POSIÇÃO | EPS | MIN_SAMPLE | NÚMERO DE CLUSTERS | SILHOUETTE |
|---|---|---|---|---|
| 1 | 0.5 | 20.0 | 2.0 | 0.634754 |
| 2 | 0.5 | 50.0 | 2.0 | 0.632289 |
| 3 | 0.5 | 100.0 | 2.0 | 0.631526 |
| 4 | 0.5 | 5.0 | 2.0 | 0.617227 |
| 5 | 0.5 | 10.0 | 2.0 | 0.608664 |
| ... | ... | ... | ... | ... |
| 15 | 0.3 | 5.0 | 4.0 | 0.475543 |

Fonte: Elaborada pelo autor (2020)

Embora a aplicação do modelo DBSCAN tenha indicado o melhor *silhouette* para dois clusters, conforme observado na Tabela 17 - Valores de *Silhouette* para DBScan, optou-se, neste caso, por selecionar o décimo quinto colocado da lista, por ser ele o primeiro da tabela com 4 clusters, número mínimo definido no critério de teste. A figura abaixo mostra o resultado após aplicação da modelagem DBScan para a aplicação dos parâmetros para 4 clusters.

Figura 17 - Plotando o *dataframe* após clusterização DBScan




Fonte: Elaborada pelo autor (2020)

### 8.3.4 Modelo Hierárquico

*Hierarquical Clutering* é um gênero para algoritmos de clusterização que constrói clusters aninhados, mesclando-os ou dividindo-os sucessivamente. Essa hierarquia de clusters é representada como uma árvore (dendrograma). A raiz da árvore é o agrupamento único que reúne todas as amostras, as folhas sendo os agrupamentos com apenas uma amostra (SCIKIT-LEARN, 2020). Possui dois tipos de estratégias:

- *Agglomerative Cluster*: cuja abordagem é "baixo para cima". Nessa abordagem, cada observação começa em seu próprio cluster e pares de clusters são mesclados à medida que se sobe na hierarquia.
- *Divisive Cluster*: cuja abordagem é "Cima para baixo". Aqui, todas as observações começam em um único cluster e as divisões são executadas recursivamente à medida que desce na hierarquia.

Neste trabalho, optou-se por utilizar a estratégia de clusterização hierárquica *Agglomerative Clustering*. Embora exija pouco mais de recursos de máquina, em alguns casos pode-se chegar a resultados melhores.

Da mesma forma, foi aplicada a análise do *Silhouette coefficient* variando o número de clusters de 4 a 25. Os melhores resultados estão listados na tabela 18, abaixo:

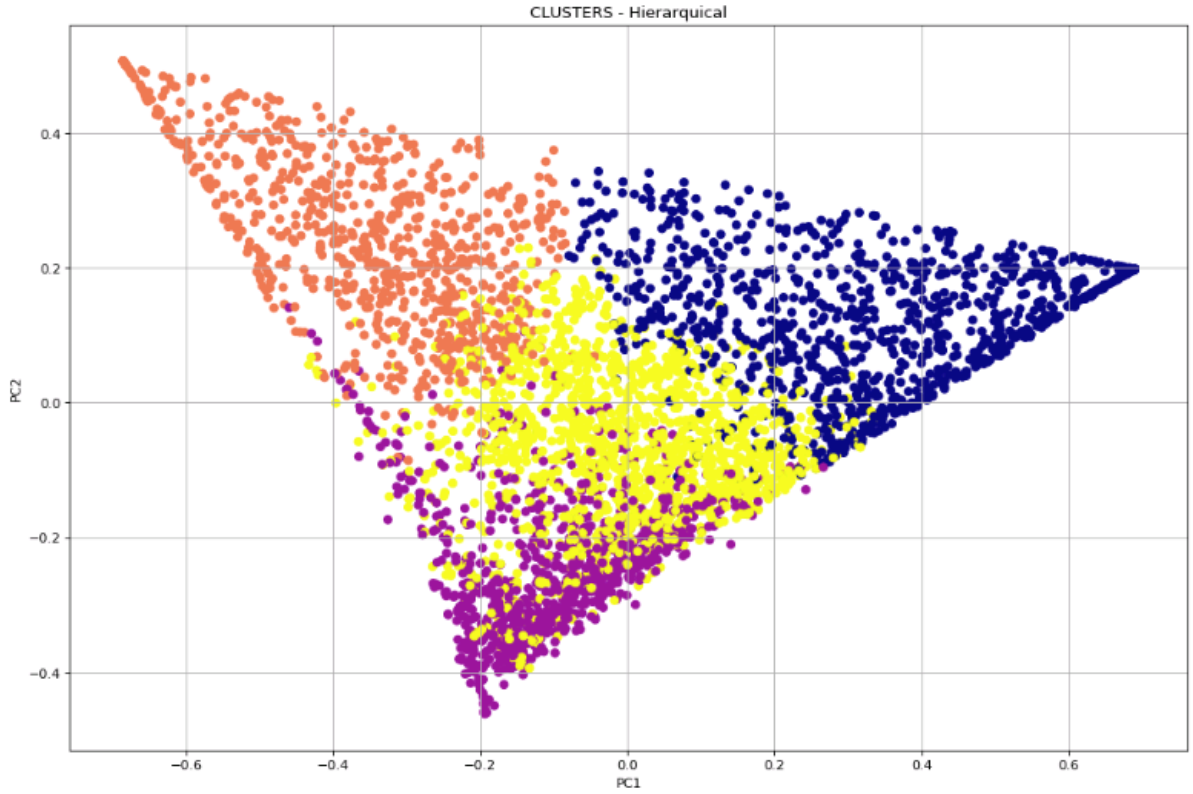
Tabela 18 - Valores de *Silhouette* para Hierárquico

| POSIÇÃO | NÚMERO DE CLUSTERS | SILHOUETTE |
|---|---|---|
| 1 | 4.0 | 0.298618 |
| 2 | 23.0 | 0.274275 |
| 3 | 24.0 | 0.272528 |
| 4 | 22.0 | 0.270560 |
| 5 | 21.0 | 0.266504 |

Fonte: Elaborada pelo autor (2020)

Os agrupamentos sugeridos pelo modelo hierárquico, considerando os valores obtidos na primeira linha da Tabela 18 - Valores de *Silhouette* para Hierárquico, indicaram o número de 4 clusters. A figura abaixo mostra o resultado após aplicação da modelagem Hierárquica.

Figura 18 - Plotando o *dataframe* após clusterização Hierárquica




Fonte: Elaborada pelo autor (2020)

## 8.4 AVALIAÇÃO DOS MODELOS

Na atividade de avaliação dos modelos, foram realizadas análises quantitativas dos resultados obtidos pelos quatro modelos aplicados. Cada modelo resultou num número de clusters diferentes, com uma quantidade de membros em cada grupo. Na tabela a seguir, estão listados os resultados nominais obtidos de cada modelagem.

Tabela 19 - Quantidade de clusters e de membros em cada cluster

| **K-MEANS** | | **MEAN-SHIFT** | | **DBSCAN** | | **HIERARQUICAL** | |
|---|---|---|---|---|---|---|---|
| **CLUSTER** | **QTD DE MEMBROS** | **CLUSTER** | **QTD DE MEMBROS** | **CLUSTER** | **QTD DE MEMBROS** | **CLUSTER** | **QTD DE MEMBROS** |
| 0 | 174 | 0 | 4359 | -1 | 65 | 0 | 1139 |
| 1 | 683 | 1 | 4 | 0 | 4288 | 1 | 984 |
| 2 | 63 | 2 | 3 | 1 | 7 | 2 | 809 |
| 3 | 473 | 3 | 2 | 2 | 10 | 3 | 1438 |
| 4 | 449 | 4 | 1 | | | | |
| 5 | 157 | 5 | 1 | | | | |
| 6 | 124 | | | | | | |
| 7 | 409 | | | | | | |
| 8 | 770 | | | | | | |

| K-MEANS | | MEAN-SHIFT | | DBSCAN | | HIERARQUICAL | |
|---|---|---|---|---|---|---|---|
| **CLUSTER** | **QTD DE MEMBROS** | **CLUSTER** | **QTD DE MEMBROS** | **CLUSTER** | **QTD DE MEMBROS** | **CLUSTER** | **QTD DE MEMBROS** |
| 9 | 225 | | | | | | |
| 10 | 843 | | | | | | |

Fonte: Elaborada pelo autor (2020)

Nota-se, analisando a Tabela 19 - Quantidade de clusters e de membros em cada cluster, que dois modelos (DBScan e Mean_Shift) não atingiram os critérios mínimos de sucesso definidos no item 5.3. METAS DE MINERAÇÃO E CRITÉRIOS DE SUCESSO, referente ao quantitativo mínimo de membros em cada grupo, ou seja, **30 membros**.

A partir desta análise, optou-se por **não realizar** as análises qualitativas nos resultados dos modelos DBScan e Mean Shift a partir deste momento do projeto. Assim sendo, foram realizadas as avaliações, nas seções seguintes, apenas dos resultados dos modelos: K-means e Hierárquico.

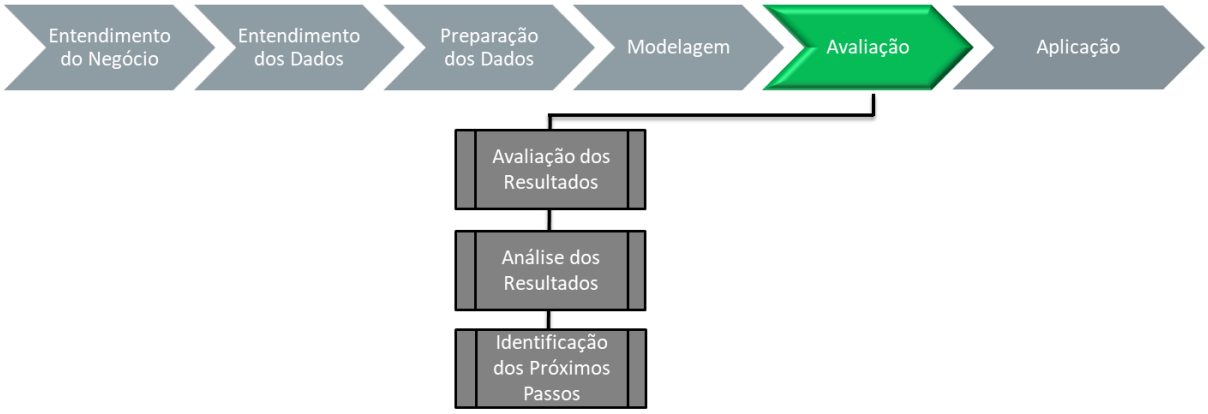
# 9 AVALIAÇÃO

Nesta etapa do *CRISP-DM*, foram realizadas avaliações qualitativas nos resultados dos modelos. Procurou-se identificar, nos resultados gerados pelos algoritmos, quais realmente respondem às questões elencadas no item 5.2.1 - Objetivos Específicos. Procurou-se responder a seguinte pergunta: Os resultados obtidos foram suficientemente eficazes para atingirem o objetivo elencado?

Para ajudar a responder esta pergunta, utilizou-se da ferramenta de exploração e visualização de dados SAS VA, disponível no TCU, que provê funções de elaboração de painéis de visualização.

As atividades executadas nessa etapa estão ilustradas na figura a seguir.

Figura 19 - Detalhamento da fase “Avaliação” do *CRISP-DM*




Fonte: Elaborada pelo autor (2020)

## 9.1 AVALIAÇÃO DOS RESULTADOS

As atividades previstas nessa etapa contemplam a terceira e última meta de mineração definida na Figura 4 – Produtos desenvolvidos no trabalho, mostrada a seguir.

Figura 20 - Dashboard


Fonte: Elaborada pelo autor (2020)

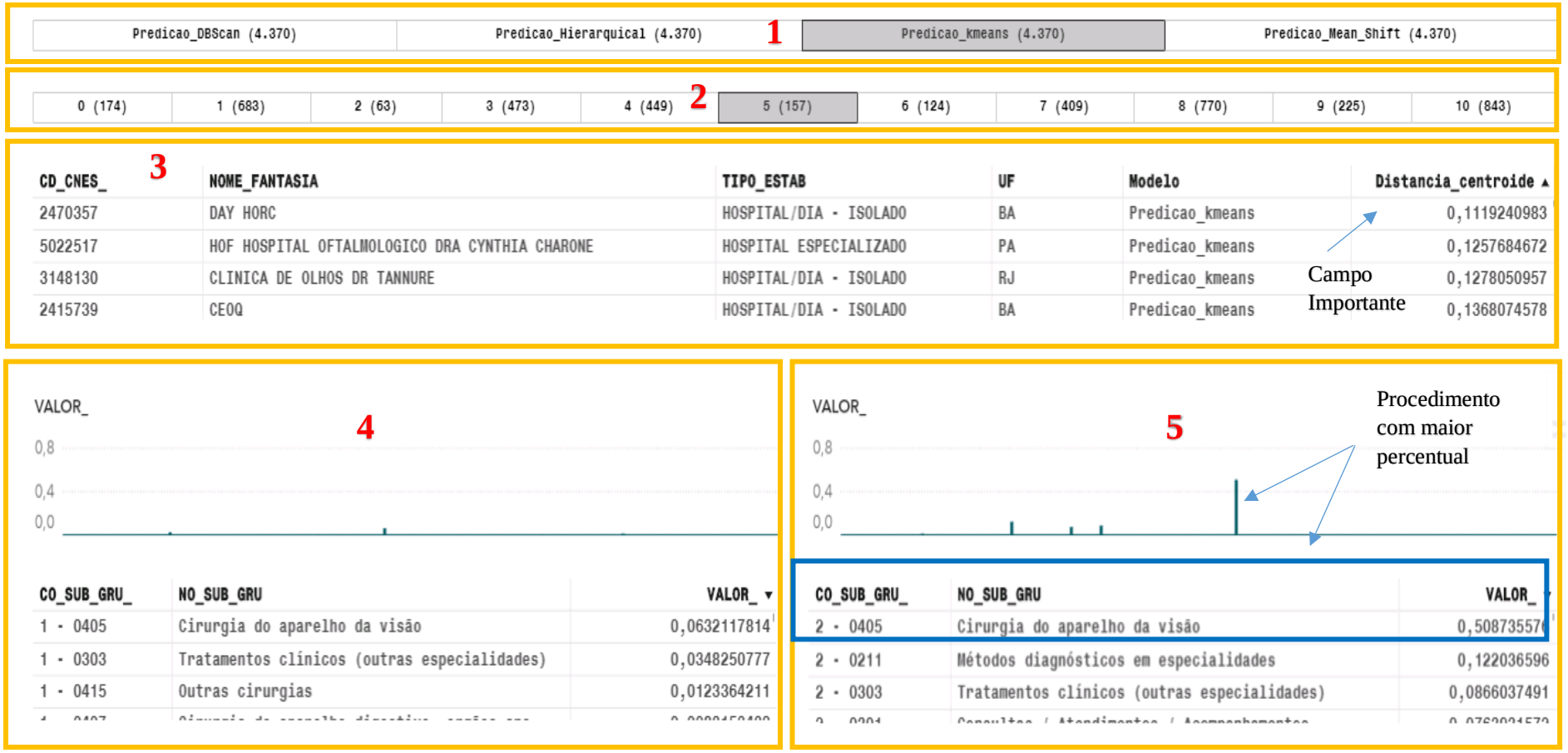
Conforme já mencionado, a avaliação dos resultados foi realizada apenas para os modelos K-means e Hierárquico, tendo em vista que os modelos DBScan e Mean Shift não atingiram os critérios mínimos quantitativos elencados nos critérios de sucesso.

Na avaliação dos resultados, tentou-se descobrir se os modelos conseguiram, de fato, agrupar unidades de saúde em subconjuntos similares, de acordo com os perfis de procedimento de cada uma. Os resultados foram feitos para os dois modelos individualmente, e estão mostrados a seguir.

### 9.1.1 Modelo K-means

Para esta análise, utilizou-se ferramenta de exploração de dados e construção de painéis SAS VA. Esta ferramenta, homologada pela SETIC, proporciona a construção de dashboards bem como a manipulação dos dados para exploração e visualização de dados. Na Figura 21 - Exemplo da análise do perfil de hospitais, abaixo, é mostrado um exemplo de como foi feita a avaliação qualitativa dos clusters e dos modelos.

Figura 21 - Exemplo da análise do perfil de hospitais




Fonte: Elaborada pelo autor (2020)

O campo 1 da Figura 21 - Exemplo da análise do perfil de hospitais, mostra o botão de seleção do tipo de algoritmo de clusterização utilizado. Nota-se que o botão Predição_Kmeans está selecionado, forçando com que os demais campos da tela sejam filtrados por esta seleção.

O campo 2 corresponde ao botão dos clusters da Predição_Kmeans. Da mesma forma, tendo em vista que o cluster 5 também encontra-se selecionado, os demais campos da tela (campos 3, 4 e 5) apresentam dados desta seleção.

O campo 3 contém as unidades de saúde (DMUs) que pertencem ao cluster selecionado (no caso, cluster 5 da Predição_Kmeans). Importante ressaltar que, dentre as colunas da lista, destaca-se a "Distância_centroide", que contém a distância da unidade de saúde ao centroide do seu cluster. Ordenando os valores desta coluna, consegue-se identificar as unidades que estão mais próximas e mais distantes do centroide do seu cluster, o que pôde facilitar demasiadamente a análise qualitativa por parte da Unidade Técnica.

Os campos 4 e 5 da figura contêm as informações do perfil de procedimentos realizados no cluster. O campo 4 apresenta informações do SIH, enquanto que o campo 5, do SIA. Os campos 4 e 5 estão divididos, cada um, em dois gráficos cada, sendo um na parte de cima (gráfico de linhas) e outro na parte de baixo (tabela de lista). Estes dois gráficos possuem essencialmente a mesma informação, sendo que o gráfico de linhas possui uma representação visual, enquanto que a tabela de lista, representação numérica. Os valores apresentados nestes dois gráficos representam os percentuais dos procedimentos realizados no cluster, que, na verdade, correspondem aos percentuais dos seus centroides.

Analisando a Figura 21 - Exemplo da análise do perfil de hospitais, nota-se que a Predição_kmeans agrupou no cluster cinco (5), num total de 157 unidades, unidades com perfil de atendimento essencialmente oftalmológico, tendo em vista que o procedimento mais frequente deste cluster é o de "Cirurgia do aparelho de visão".

Na Tabela 20 - K-means - Lista do perfil de hospitais, abaixo, foram realizadas as mesmas análises para os demais clusters da Predição_kmeans. Nesta tabela, foram incluídos os procedimentos cujos percentuais somados chegaram à, pelo menos, 70% do total no cluster. Na coluna "Perfil" desta tabela, contém uma denominação (representação) considerada a mais comum do clusters, indicando, possivelmente, seu perfil de atendimento.

Tabela 20 - K-means - Lista do perfil de hospitais

| CLUSTER K-MEANS | QTD UNID | CO_SUB_GRU | NO_SUB_GRUPO | VALOR_ | Perfil |
|---|---|---|---|---|---|
| 0 | 174 | 2 - 0202 | Diagnóstico em laboratório clínico | 0,415825 | Diagnóstico em laboratório clínico e Consultas / Atendimentos / Acompanhamentos |
| | | 2 - 0301 | Consultas / Atendimentos / Acompanhamentos | 0,1587637 | |
| | | 1 - 0303 | Tratamentos clínicos (outras especialidades) | 0,1464029 | |
| 1 | 683 | 1 - 0303 | Tratamentos clínicos (outras especialidades) | 0,5631327 | Tratamentos clínicos (outras especialidades) |
| | | 2 - 0301 | Consultas / Atendimentos / Acompanhamentos | 0,1460706 | |
| 2 | 63 | 2 - 0305 | Tratamento em nefrologia | 0,5699535 | Nefrologia |
| | | 1 - 0408 | Cirurgia do sistema osteomuscular | 0,1297837 | |
| 3 | 473 | 1 - 0406 | Cirurgia do aparelho circulatório | 0,121558 | Generalista |
| | | 1 - 0303 | Tratamentos clínicos (outras especialidades) | 0,0994936 | |
| | | 1 - 0415 | Outras cirurgias | 0,0850764 | |
| | | 1 - 0408 | Cirurgia do sistema osteomuscular | 0,0628884 | |
| | | 2 - 0301 | Consultas / Atendimentos / Acompanhamentos | 0,0533821 | |
| | | 1 - 0407 | Cirurgia do aparelho digestivo, orgãos anexos e parede abdominal | 0,0549579 | |
| | | 2 - 0204 | Diagnóstico por radiologia | 0,0355525 | |
| | | 1 - 0409 | Cirurgia do aparelho geniturinário | 0,0303377 | |
| | | 2 - 0304 | Tratamento em oncologia | 0,029357 | |
| | | 2 - 0211 | Métodos diagnósticos em especialidades | 0,0292641 | |
| | | 1 - 0505 | Transplante de orgãos, tecidos e células | 0,0291486 | |
| | | 2 - 0303 | Tratamentos clínicos (outras especialidades) | 0,0282313 | |

| CLUSTER K-MEANS | QTD UNID | CO_SUB_GRU | NO_SUB_GRUPO | VALOR_ | Perfil |
|---|---|---|---|---|---|
| | | 2 - 0202 | Diagnóstico em laboratório clínico | 0,0249583 | |
| | | 2 - 0305 | Tratamento em nefrologia | 0,023193 | |
| 4 | 449 | 1 - 0303 | Tratamentos clínicos (outras especialidades) | 0,878539 | Tratamentos clínicos (outras especialidades) |
| 5 | 157 | 2 - 0405 | Cirurgia do aparelho da visão | 0,5087356 | Oftalmológico |
| | | 2 - 0211 | Métodos diagnósticos em especialidades | 0,1220366 | |
| | | 2 - 0303 | Tratamentos clínicos (outras especialidades) | 0,0866037 | |
| 6 | 124 | 2 - 0304 | Tratamento em oncologia | 0,3806854 | Oncologia |
| | | 1 - 0415 | Outras cirurgias | 0,0932587 | |
| | | 1 - 0303 | Tratamentos clínicos (outras especialidades) | 0,0820738 | |
| | | 1 - 0416 | Cirurgia em oncologia | 0,0750439 | |
| | | 1 - 0406 | Cirurgia do aparelho circulatório | 0,0433905 | |
| | | 2 - 0305 | Tratamento em nefrologia | 0,0332205 | |
| 7 | 409 | 2 - 0301 | Consultas / Atendimentos / Acompanhamentos | 0,7414844 | Consultas / Atendimentos / Acompanhamentos |
| 8 | 770 | 2 - 0301 | Consultas / Atendimentos / Acompanhamentos | 0,4007916 | Consultas / Atendimentos / Acompanhamentos e Tratamentos clínicos (outras especialidades) |
| | | 1 - 0303 | Tratamentos clínicos (outras especialidades) | 0,224103 | |
| | | 2 - 0202 | Diagnóstico em laboratório clínico | 0,0653636 | |
| | | 1 - 0310 | Parto e nascimento | 0,0428474 | |
| 9 | 225 | 1 - 0310 | Parto e nascimento | 0,2863888 | Maternidades |
| | | 1 - 0411 | Cirurgia obstétrica | 0,2812492 | |
| | | 1 - 0303 | Tratamentos clínicos (outras especialidades) | 0,1540771 | |
| 10 | 843 | 1 - 0303 | Tratamentos clínicos (outras especialidades) | 0,3249656 | Tratamentos clínicos (outras especialidades), Consultas / |

| CLUSTER K-MEANS | QTD UNID | CO_SUB_GRU | NO_SUB_GRUPO | VALOR_ | Perfil |
|---|---|---|---|---|---|
| | | 1 - 0301 | Consultas / Atendimentos / Acompanhamentos | 0,1146482 | Atendimentos / Acompanhamentos e Cirurgia obstétrica |
| | | 1 - 0411 | Cirurgia obstétrica | 0,0751928 | |
| | | 1 - 0407 | Cirurgia do aparelho digestivo, orgãos anexos e parede abdominal | 0,0609386 | |
| | | 1 - 0310 | Parto e nascimento | 0,0575289 | |
| | | 1 - 0408 | Cirurgia do sistema osteomuscular | 0,0368213 | |
| | | 2 - 0204 | Diagnóstico por radiologia | 0,0330018 | |

Fonte: Elaborada pelo autor (2020)

### 9.1.2 Modelo Hierárquico

Na Tabela 21 - Hierárquico - Lista do perfl de hospitais, é mostrado o resultado da mesma análise realizada na tabela anterior, só que para o modelo Hierárquico. Os seguintes perfis foram identificados.

Tabela 21 - Hierárquico - Lista do perfl de hospitais

| CLUSTER HIERARQ | QTD CNES | CO_SUB_GRU | NO_SUB_GRUPO | VALOR_ | Perfil |
|---|---|---|---|---|---|
| 0 | 1139 | 1 - 0303 | Tratamentos clínicos (outras especialidades) | 0,68044151 | Tratamentos clínicos (outras especialidades) |
| | | 2 - 0301 | Consultas / Atendimentos / Acompanhamentos | 0,11991991 | |
| 1 | 984 | 1 - 0303 | Tratamentos clínicos (outras especialidades) | 0,11287862 | Generalista |
| | | 2 - 0405 | Cirurgia do aparelho da visão | 0,10054473 | |
| | | 2 - 0301 | Consultas / Atendimentos / Acompanhamentos | 0,06991201 | |
| | | 2 - 0304 | Tratamento em oncologia | 0,06854068 | |
| | | 1 - 0406 | Cirurgia do aparelho circulatório | 0,06792511 | |
| | | 1 - 0415 | Outras cirurgias | 0,05846029 | |
| | | 1 - 0305 | Tratamento em nefrologia | 0,05813644 | |
| | | 1 - 0407 | Cirurgia do aparelho digestivo, orgãos anexos e parede abdominal | 0,04418051 | |
| | | 2 - 0211 | Métodos diagnósticos em especialidades | 0,03866005 | |
| | | 1 - 0408 | Cirurgia do sistema osteomuscular | 0,03313841 | |

| | | | | | |
|---|---|---|---|---|---|
| | | 1 - 0303 | Tratamentos clínicos (outras especialidades) | 0,03030566 | |
| | | 2 - 0202 | Diagnóstico em laboratório clínico | 0,02990632 | |
| 2 | 809 | 2 - 0301 | Consultas / Atendimentos / Acompanhamentos | 0,60203018 | Consultas / Atendimentos / Acompanhamentos |
| | | 1 - 0303 | Tratamentos clínicos (outras especialidades) | 0,14292569 | |
| 3 | 1438 | 1 - 0303 | Tratamentos clínicos (outras especialidades) | 0,27277056 | Tratamentos clínicos (outras especialidades), Diagnóstico em laboratório clínico e Cirurgia obstétrica |
| | | 2 - 0202 | Diagnóstico em laboratório clínico | 0,15420023 | |
| | | 1 - 0411 | Cirurgia obstétrica | 0,10296885 | |
| | | 2 - 0202 | Diagnóstico em laboratório clínico | 0,09231686 | |
| | | 1 - 0310 | Parto e nascimento | 0,0937303 | |

Fonte: Elaborada pelo autor (2020)

## 9.2 ANÁLISE DOS RESULTADOS

Nessa etapa, foram analisados os resultados dos agrupamentos realizados considerando o enfoque do cliente, confrontando os resultados obtidos nos algoritmos com os critérios de sucesso de negócio definidos.

Analisando os grupos gerados pelos dois modelos (K-means e Hierarquico), observa-se que o modelo K-means resultou no agrupamento maior de clusters, e com perfil mais especializado do que o agrupamento do modelo sugerido pelo Hierárquico.

No resultado do K-means, observa-se a existência de, pelo menos, quatro clusters com perfis bastante específicos. Na tabela abaixo estes clusters são listados, e nota-se que os procedimentos mais frequentes dos clusters, contabilizados em percentuais, ultrapassam 45% do total, demonstrando a existência clara de homogeneidade nestes grupos formados pelo modelo K-means.

Tabela 22 - Clusters do K-means com perfis específicos

| NUMERO DO CLUSTER | PROCEDIMENTOS MAIS FREQUENTES | PERFIL | % DOS PROCEDIMENTOS MAIS FREQUENTES |
|---|---|---|---|
| 2 | - Tratamento em nefrologia | Nefrologia | 57% |
| 5 | - Cirurgia do aparelho da visão | Oftalmologia | 50% |
| 6 | - Tratamento em oncologia<br>- Cirurgia em oncologia | Oncologia | 45% |
| 9 | - Parto e nascimento<br>- Cirurgia obstétrica | Maternidade | 56% |

Fonte: Elaborada pelo autor (2020)

Para os demais clusters do modelo K-means, nota-se que o agrupamento se deu mais pelas características de abrangência e generalidade do que pela especificidade. Um exemplo disso é o cluster 3, onde se identifica uma gama grande de procedimentos realizados com percentuais pequenos de cada um, motivo polo qual recebeu a denominação de “generalista”. Há de se destacar, entretanto, que este perfil de estabelecimento de saúde de fato existe, e estas unidades devem ser agrupadas separadamente.

Quanto ao modelo Hierárquico, o algoritmo agrupou conjunto de clusters mais diversos, com características menos específicas e mais heterogêneas. Em nenhum dos clusters sugeridos pelo modelo Hierárquico, encontra-se grupos com perfil especialista, como encontrado no modelo K-means. Nota-se que os procedimentos “Cirurgia do aparelho circulatório”, “Tratamento em nefrologia” e “Tratamento em oncologia”, que no modelo K-means estavam separados em clusters diferentes, foram agrupados todos no cluster 1 do modelo Hierárquico.

Há de se destacar, entretanto, que ambos os modelos apresentaram alguns resultados similares. O cluster 0 do Hierárquico, por exemplo, com 1139 membros, possui como principal procedimento o “Tratamentos clínicos - outras especialidades”. O modelo K-means realizou agrupamento parecido, porém dividido em dois clusters diferentes (1 e 4), que somados totalizam 1132 unidades, bem próximo do modelo Hierárquico. Esse resultado indica a existência de um tipo de representação com características muito similares, relacionadas ao perfil de procedimento “Tratamentos clínicos - outras especialidades”.

Cumpre destacar que os resultados obtidos foram apresentados ao cliente, SecexSaúde, que sinalizou a utilização do modelo k-means como o mais indicado na utilização na análise DEA, tendo em vista apresentar conjuntos com maior especificidade, homogeneidade e semelhança entre as DMUs do mesmo cluster.

## 9.3 DEFINIÇÃO DOS PRÓXIMOS PASSOS

Como próximos passos, visando o melhoramento do resultado deste projeto bem como proposição de trabalhos futuros, sugerem-se algumas atividades, das quais se destacam:

1) Realização de nova rodada de clusterização apenas nos clusters com grandes quantidades de membros e perfil heterogêneo. No caso do modelo k-means, por exemplo, o cluster 8 (com 770 unidades) pode conter perfis ainda não identificados, que poderão eventualmente ser desmembrados em subconjuntos menores. O mesmo pode ser feito para os clusters 1 e 3 do modelo Hierárquico.
2) Aumentar a granularidade do agrupamento usado no código SIGTAP (de 4 dígitos para 6 dígitos), utilizando a FORMA do código, ao invés do

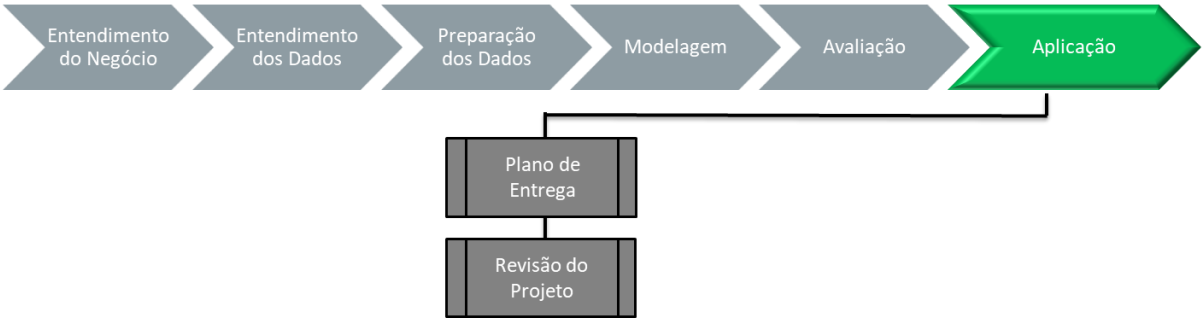
SUBGRUPO. Embora possa aumentar demasiadamente o tempo de processamento, bem como dificultar a visualização dos perfis no painel, seu resultado pode trazer resultados bastante interessantes.

3) Utilizar dados de procedimentos realizados (SIH e SIA) de período diferente do utilizado neste trabalho. Uma vez que já estão disponíveis no Labcontas, selecionar dados de anos diferentes não parece ser muito complexo, e pode agregar informações úteis como: confirmação dos resultados obtidos neste projeto, ou ainda a revelação de eventuais mudanças de perfis de unidades de saúde no decorrer dos anos.
4) Realizar Clusterizações de outros tipos de unidades de saúde, a exemplo de unidades geridas apenas por Organizações Sociais ou entidades geridas apenas por Municípios.

# 10 APLICAÇÃO

Esta fase do *CRISP-DM* abrange as atividades realizadas para implantação e sustentação da solução. Pretende-se, nesta etapa, disponibilizar os produtos desenvolvidos durante a execução deste trabalho de forma a serem utilizados pelo cliente nos trabalhos futuros. As atividades previstas nessa etapa estão mostradas na Figura 22 - Detalhamento da fase "Aplicação" do *CRISP-DM*, abaixo.

Figura 22 - Detalhamento da fase "Aplicação" do *CRISP-DM*




Fonte: Elaborada pelo autor (2020)

## 10.1 PLANO DE ENTREGA

O plano de entrega compreende definir o que se pretende fazer com os resultados alcançados. Almeja-se que os produtos desenvolvidos no decorrer do trabalho, listados na Figura 4 – Produtos desenvolvidos no trabalho, sejam implantados no ambiente de produção do Tribunal, mais especificamente pela Unidade Técnica em questão, em forma de serviços. Os produtos a serem entregues são:

a) **Internalização dos dados**. Pretende-se que os algoritmos implementados em caderno *Jupyter Notebook*, empregando linguagem *Python* (PYTHON, 2001) e a biblioteca Pandas (PROJECT, 2020) sejam disponibilizados no servidor *srv-rstudio* e sua execução agendada de forma automática e periódica.

b) **Clusterização**. Pretende-se disponibilizar os algoritmos implementados em caderno *Jupyter Notebook*, empregando linguagem *Python* (PYTHON, 2001), usando as bibliotecas *Pandas* (PROJECT, 2020) e *Pysus* (COELHO, 2018), em área específica a ser definida pelo Núcleo de Dados da SecexSaúde, de forma que possam ser executados na medida em que surgirem novas demandas de análises específicas.

c) **Dashboard**. Pretende-se disponibilizar painel utilizado com acessos aos dados internalizados e clusterizados, com possibilidade de acesso interno e externo ao TCU.

Embora os produtos estejam prontos, os mesmos requerem homologação por parte da SecexSaúde. Atualmente a Unidade Técnica está em processo de avaliação e homologação dos mesmos. Tão logo sejam testados e validados, os produtos deste trabalho serão disponibilizados e colocados em produção, obedecendo critérios de classificação da informação tanto no que tange à confidencialidade da informação, estabelecidos pela SecexSaúde, como também por legislação pertinente.

## 10.2 REVISÃO DO PROJETO

Este documento contém os passos utilizados no decorrer do trabalho. É, portanto, a espinha dorsal de conhecimento do projeto. Os demais produtos desenvolvidos, juntamente com a apresentação feita à banca examinadora, estarão disponibilizados para consulta em base específica e em momento oportuno.

# 11 CONSIDERAÇÕES FINAIS

O presente trabalho teve como objetivo a construção de um modelo de mineração capaz de realizar a clusterização das unidades de saúde cadastradas no sistema CNES. Durante a execução do trabalho, identificou-se que seria necessário realizar tarefas adicionais como: a internalização dos dados do sistema CNES no datacenter do TCU e a construção de painéis de visualização destes dados através de *dashboards*. Estes objetivos revelaram a necessidade cada vez mais imperiosa do TCU, mais especificamente da SecexSaúde, de utilizar recursos de tecnologia da informação e de ciência de dados nos trabalhos voltados ao Controle.

Os resultados obtidos demonstram que os objetivos foram atingidos. Foram desenvolvidos dois algoritmos, utilizando linguagem *Python,* para internalização dos dados do sistema CNES no ambiente computacional do TCU. A carga inicial dos dados, de 2012 a 2019, foi concluída, e o algoritmo para manter os dados atualizados está pronto. Atualmente, o TCU pode realizar uma ampla quantidade de novas consultas, análises e cruzamentos de dados relacionados à saúde, até então não possíveis ao Tribunal, ampliando o leque de possibilidades de auditorias e achados.

Os modelos de mineração desenvolvidos atenderam aos critérios de sucesso definido. Os resultados apresentados comprovaram sua eficácia no reconhecimento de padrões por vias computacionais, distinguindo perfis de hospitais por tipo de procedimento realizado. Foram empregados quatro modelos de mineração, sendo que o modelo considerado mais vantajoso foi o K-means. Este modelo agrupou em 11 clusters os estabelecimentos de saúde com características distintas de hospitais no Brasil. Constatou-se uma clara e incontestável distinção entre alguns clusters agrupados pelo modelo, como os hospitais que tratam de aparelho da visão, por exemplo. Outros clusters foram agrupados mais pelas características de abrangência do que pelos perfis de especialidades, indicando serem unidades de saúde com atributos de maior heterogeneidade.

A criação de painéis revelou uma poderosa ferramenta para análise e exploração dos dados. Este instrumento possibilitou, além da visualização crua dos dados internalizados, o exame dos grupos formados nas clusterizações, possibilitando ajustes nos rumos do trabalho na medida em que permitia análise exploratórias dos resultados de forma fácil, amigável e manipulável.

Sobre o trabalho em si, os resultados foram excelentes. Mesmo considerando o pouco tempo para realização das atividades e os percalços encontrados durante algumas etapas, tudo

foi superado. Constatou-se que houve uma grande contribuição à unidade técnica para o alcance de seus objetivos.

O *feedback* recebido pela Área de Negócio revelou a necessidade de continuidade do presente trabalho. Além de possíveis aprimoramentos, há também inúmeras novas proposições de agrupamentos que poderão ser desmembrados utilizando outras variáveis e enfoques. Como sugestões, foram elencados alguns trabalhos ou proposições para projetos futuros, bem como o próprio refinamento deste. Vislumbrou-se que o ganho em experiência e conhecimento adquiridos por todos os envolvidos no trabalho foi bastante proveitoso, sendo incontroverso a certeza de alcance do alvo maior do trabalho de conclusão de curso: demonstrar o valor da utilização de recursos de TI e da análise de dados a favor dos processos de negócios, em particular do Controle.

## 12 BIBLIOGRAFIA

BM, U. **ANÁLISE DE EFICIENCIA NA ATENÇÃO HOSPITALAR - RASCUNHO**. BANCO MUNDIAL E UFRJ. RIO DE JANEIRO. 2019.

BRASIL. **Constituição da República Federativa do Brasil, promulgada em 05 de outubro de 1988**. Brasília: [s.n.], 1988.

BRASIL. **Lei No 8.080 19 de setembro de**. Brasília: que dispõe sobre condições para a promoção, proteção e recuperação da saúde, sua organização e seu funcionamento, 1990.

BRASIL. **Lei no 12.527, de 18 de novembro de 2011**. Brasília: que dispõe sobre os procedimentos a serem observados pela União, Estados, Distrito Federal e Municípios para garantir o acesso à informação., 2011.

CHENG, Y. **Mean Shift, Mode Seeking, and Clustering**. [S.l.]: IEEE - Transactions on Pattern Analysis and Machine Intelligence, 1995.

COELHO, F. C. PySUS Documentation, p. 23, Dez 2018. Disponivel em: <https://buildmedia.readthedocs.org/media/pdf/pysus/latest/pysus.pdf>. Acesso em: 15 mar. 2020.

DATASCIENCE. **Portal Data Science**, 2020. Disponivel em: <https://portaldatascience.com/introducao-a-clusterizacao-e-os-diferentes-metodos/>. Acesso em: 10 fev. 2020.

DATASUS. **CNES - Informe Técnico 201-06 - DADOS AUXILIARES**, Brasília, 2017. Disponivel em: <ftp://ftp.datasus.gov.br/dissemin/publicos/CNES/200508_/Auxiliar/>. Acesso em: 4 out. 2019.

DEVELOPERS, T. S.-Y. **Elbow Method**, 2016. ISSN Revision 682b3528. Disponivel em: <https://www.scikit-yb.org/en/latest/api/cluster/elbow.html>. Acesso em: 26 fev. 2020.

FNS. **FUNDO NACIONAL DA SAÚDE**, 2019. Disponivel em: <https://consultafns.saude.gov.br/#/consolidada>. Acesso em: 2020 fev 26.

GONÇALVES, A. C.; NORONHA, C. P. Avaliando a eficiência dos hospitais gerais do SUS, através da metodologia da Análise Envoltória De Dados – Dea. **ACADEMUS REVISTA CIENTÍFICA DA SAÚDE**, Rio de Janeiro, v. 1, fev. 2002.

HAN J, K. M. A. T. **Spatial clustering methods in data mining**. [S.l.]: [s.n.], 2001.

IBM. CRISP-DM (Cross-Industry Standard Process for Data Mining) Guide. **IBM® SPSS® Modeler**, 2014. Disponivel em: <https://www.ibm.com/support/knowledgecenter/en/SS3RA7_15.0.0/com.ibm.spss.crispdm.help/crisp_overview.htm>. Acesso em: 10 fev. 2020.

KUANG HONGYU, V. L. M. S. G. J. D. O. J. Análise de Componentes Principais: resumo teórico, aplicação e interpretação. **Engineering and Science**, v. 1, n. 5, p. 83-89, 2015. Disponivel em: <http://papers.nips.cc/paper/2628-a-direct-formulation-for-sparse-pca-using-semidefinite-programming.pdf>.

MAXWEL. Clusterização de Dados. **PUC-Rio**, Rio de Janeiro, 2014.

MEDIUM, 2020. Disponivel em: <https://medium.com/@kvmoura/crisp-dm-79580b0d3ac4>. Acesso em: 11 mar. 2020.

MEDIUM. Neuronio.ai. **Aprendizado não supervisionado com K-means**, 2020. Disponivel em: <https://medium.com/neuronio-br/aprendizado-n%C3%A3o-supervisionado-com-k-means-f4272dee98a0>. Acesso em: 07 abr. 2020.

OCDE. Hospital Performance - Organização para o Comercio e Desenvolvimento Econômico. **Hospital Performance**, 2020. Disponivel em: <https://www.oecd.org/health/health-systems/hospital-performance.htm>. Acesso em: 24 fev 2020.

PROJECT, T. P. PANDAS. **Pandas - Python Data Analysis Library**, 2020. Disponivel em: <https://pandas.pydata.org>. Acesso em: 10 mar. 2020.

PYTHON, S. F. PYTHON.ORG. **PYTHON**, 2001. Disponivel em:<. Acesso em: 10 mar. 2020.

SAÚDE, M. D. **PORTARIA 376**. Brasília: [s.n.], 2000.

SAÚDE, M. D. **PORTARIA MS/SAS 511**. Brasília: [s.n.], 2000.

SAÚDE, M. D. CNES. **Institui o Cadastro Nacional de Estabelecimentos de Saúde**, Brasília, 2015.

SAÚDE, M. D. Portaria nº 1646. **Portaria nº 1646**, 2015. Disponivel em: <http://bvsms.saude.gov.br/bvs/saudelegis/gm/2017/MatrizesConsolidacao/comum/13357.html>. Acesso em: 2020 mar. 19.

SAÚDE, M. D. Cadastro Nacional de Estabelecimentos de Saúde. **DATASUS**, 2019. Disponivel em: <http://cnes.datasus.gov.br>. Acesso em: 10 set 2019.

SAÚDE, M. D. **Sistema Único de Saúde (SUS):** estrutura, princípios e como funciona, 2020. Disponivel em: <https://www.saude.gov.br/sistema-unico-de-saude>. Acesso em: 12 mar. 2020.

SCIENCE, T. D. https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1. **Towards data science**, 2016. Disponivel em: <https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1>. Acesso em: 28 fev. 2020.

SCIKIT-LEARN. **SCIKIT-LEARN**, 2019. Disponivel em: <https://scikit-learn.org/stable/modules/generated/sklearn.cluster.DBSCAN.html>. Acesso em: 25 out. 10.

SCIKIT-LEARN. CLUSTERING. **SCIKIT LEARN**, 2020. Disponivel em: <https://scikit-learn.org/stable/modules/clustering.html#mean-shift>. Acesso em: 28 fev. 2020.

SCIKIT-LEARN. SCIKIT-LEARN. **SCIKIT-LEARN**, 2020. Disponivel em: <https://scikit-learn.org/stable/modules/clustering.html#hierarchical-clustering>. Acesso em: 2020 fev. 05.

SIGTAP. https://wiki.saude.gov.br/sigtap/index.php/P%C3%A1gina_principal. **SIGTAP**, 2020. Disponivel em: <https://wiki.saude.gov.br/sigtap/index.php/P%C3%A1gina_principal>. Acesso em: 10 fev. 2020.

SOUZA, P. C. D.; SCATENA, J. H. G.; KEHRIG, R. T. Aplicação da Análise Envoltória de Dados para avaliar a eficiência de hospitais do SUS em Mato Grosso. **Physis:** Revista de Saude Coletiva, mar. 2016. Disponivel em: <https://www.scielosp.org/article/physis/2016.v26n1/289-308/pt/>. Acesso em: 11 mar. 2020.

TCU. **RELATÓRIO DE AUDITORIA**. Brasilia: [s.n.], v. TC 018.584/2014-4, 2014.

TCU. **Plano Estratégigo 2015 - 2021**. Brasília: [s.n.], 2015.

TCU. Levantamento da eficiência das unidades prestadoras de serviço de saúde de média e alta complexidades. In: TCU **TC 015.993/2019-1**. [S.l.]: [s.n.], 2019.

WILLIAM W. COOPER, L. M. S. J. Z. **Handbook on Data Envelopment Analysis**. second. ed. [S.l.]: [s.n.], v. 164, 2011.