PONTIFÍCIA UNIVERSIDADE CATÓLICA DE MINAS GERAIS
NÚCLEO DE EDUCAÇÃO A DISTÂNCIA
Pós-graduação *Lato Sensu* em Ciência de Dados e Big Data

NAHARI DE FARIA MARCOS TERENA

COMPARAÇÃO DE AGRUPAMENTO DA MORTALIDADE INFANTIL ENTRE INDÍGENAS E NÃO-INDÍGENAS EM 2019

Belo Horizonte
2023

NAHARI DE FARIA MARCOS TERENA

COMPARAÇÃO DE AGRUPAMENTO DA MORTALIDADE INFANTIL ENTRE INDÍGENAS E NÃO-INDÍGENAS EM 2019

Trabalho de Conclusão de Curso apresentado ao Curso de Especialização em Ciência de Dados e Big Data como requisito parcial à obtenção do título de especialista.

Belo Horizonte
2023

SUMÁRIO

1. Introdução

1.1. Contextualização

Atualmente e cada vez mais, técnicas computacionais são utilizadas para tratar, manipular, aplicar e gerar informações baseadas em dados nos mais diversos setores. A variedade de informações coletadas e passíveis de cruzamentos, a velocidade com que se deve gerar informações para aproveitar oportunidades e o volume de dados corroboram o uso mais frequente de recursos tecnológicos.

O crescimento da quantidade e complexidade dos dados têm também gerado alguns desafios em relação à escolha da metodologia estatística. O Aprendizado de Máquina (*Machine Learning* - ML) é reconhecido como um método promissor para apoiar o diagnóstico ou prever resultados clínicos [1].

De acordo com estimativas da ONU, de 2000 a 2017, parte das complicações que se desenvolvem durante a gravidez poderia ser tratáveis, monitoradas ou evitáveis [2].

A mortalidade infantil é, segundo a Organização Mundial de Saúde (OMS) [3], o óbito da criança ocorrido antes de completar um ano de idade. Os Objetivos de Desenvolvimento Sustentável (ODS) foram lançados em 2015 e estabelecem uma série de metas a serem atingidas no período de 15 anos, ou seja, até 2030. Entre os 17 ODS, os fatores diretos relacionados à saúde estão incluídos no ODS 3: Assegurar uma vida saudável e promover o bem-estar para todas e todos, em todas as idades.

A mortalidade infantil é um importante indicador de saúde e condições de vida de uma população. Com o cálculo da sua taxa, estima-se o risco de um nascido vivo morrer antes de chegar a um ano de vida. Valores elevados refletem precárias condições de vida e saúde e baixo nível de desenvolvimento social e econômico.

Com o objetivo de apresentar uma técnica de *Machine Learning* (ML) para endossar o aperfeiçoamento das informações acerca do óbito infantil, este relatório apresenta informações do *linkage* entre o Sistema de Informação de Mortalidade (SIM) e o Sistema de Informações sobre Nascidos Vivos (Sinasc) referente a mortalidade infantil para o ano de 2019, comparando modelos que identifiquem agrupamentos significativos.

1.2. O problema proposto

No Brasil, houve uma importante redução na mortalidade infantil ao longo das últimas décadas, devido à queda da fecundidade, à expansão do saneamento básico, à reorganização do modelo de atenção à saúde, a melhorias na atenção à saúde da criança, ao aumento na cobertura das campanhas de vacinação e na prevalência do aleitamento materno, que influenciaram a redução de doenças infecciosas nos primeiros anos de vida [4]. Constata-se ainda, as desigualdades regionais e as iniquidades relacionadas a grupos sociais considerados vulneráveis constituem grandes desafios em nosso país. Identificar padrões e falhas na assistência é necessário para a melhoria do cuidado materno infantil.

Os dados de 2019, bem como outros anos, são disponibilizados pelo Departamento de Análise Epidemiológica e Vigilância de Doenças Não Transmissíveis (DAENT/SVS/MS) da Coordenação-Geral de Informações e Análises Epidemiológicas (CGIAE) do Ministério da Saúde do Brasil.

Nos dados analisados, existem distinção significativa entre o grupo de indígenas e não-indígenas? É possível identificar padrões nos óbitos infantis? A partir dos resultados, subsidiar processos de planejamento, gestão e avaliação de políticas públicas e ações de saúde voltadas para atenção ao pré-natal, parto e a proteção de saúde infantil.

O agrupamento em si é interessante, pois descreve traços centrais no padrão de mortalidade. Podemos comparar diferentes sociedades examinando o nível de agrupamento e, dessa forma, obter informações importantes sobre a distribuição da mortalidade.

O presente trabalho não tem vínculo oficial junto aos referidos órgãos.

1.3. Objetivos

O enfoque do trabalho é baseado em três âmbitos:

- Testar a hipótese de que há independência entre o falecido ser indígena ou não ser e o grupo etário infantil de quando ocorreu o óbito.

- Identificar padrões nos respectivos grupos através de comparação dos modelos aplicados.

- Comparar as características em cada grupo em seu contexto.

2. Coleta de Dados

O levantamento de informações para este trabalho considerou dois bancos de dados, o Sistema de Informações sobre Mortalidade (SIM) com o Sistema de Informações sobre Nascidos Vivos (Sinasc) para verificar casos de mortalidade infantil em 2019.

O Sistema de Informações sobre Mortalidade (SIM) é um recurso para obtenção regular de dados sobre mortalidade no país. Com isso, as diversas esferas dos sistemas públicos ou privado de saúde, desde o nível municipal ao nacional, podem realizar e acompanhar situações, planejar e avaliar ações para a área. Os dados coletados por meio da Declaração de Óbito (DO) são digitados no SIM pelas Secretarias Municipais de Saúde ou pela Secretaria de Estado de Saúde que os encaminha para a Secretaria de Vigilância em Saúde (SVS).

O Sistema de Informações sobre Nascidos Vivos (Sinasc) foi desenvolvido para reunir informações epidemiológicas referentes aos nascimentos informados em todo território nacional. Assim, é possível subsidiar as ações relacionadas à saúde da mulher e da criança para todos os níveis do Sistema Único de Saúde (SUS); e igualmente planejar ações de atenção à gestante e ao recém-nascido.

Para realização, foi necessário agendamento na sala segura do Ministério da Saúde, de acordo com a Lei de Acesso à Informação. Depois do *linkage* entre as bases, ocorreram a anonimização e tratamentos de segurança para que não houvesse dados sensíveis passíveis de identificação.

Utilizando-se o pareamento determinístico, os campos de “Nome”, “Sobrenome” e “Município de Residência” fazem a correspondência exata entre as duas bases. No entanto, se há algum erro de digitação e/ou confusão no preenchimento de algum campo, o pareamento é reclinado. No caso do pareamento probabilístico, o valor critério de corte a ser considerado é de 0,7. De acordo com o modelo desenvolvido por Fellegi e Sunter [5], os pares podem ser classificados como os que têm correspondência exata, as possíveis correspondências e os que não são correspondentes, baseados no cálculo do score do pareamento e da regra de decisão, o valor de corte.

Por exemplo, tudo mais correspondente, “Maria Silva” e “Maria Silba” são a mesma pessoa, mas no pareamento determinístico não seria correspondente, e no probabilístico, haveria possibilidade de acordo com seu score de ser um par.

Nem todos os registros com mesmo “Nome”, “Sobrenome” e “Município de residência” apresentaram, no entanto, concordância exata nas variáveis levantadas, o que motivou o uso de métodos probabilísticos.

Os campos relativos a município eram preenchidos com os campos de acordo com o critério do IBGE. No entanto, em alguns casos, foram percebidos correspondência nos demais campos e no “Município de Residência” o preenchimento apenas dos dois primeiros dígitos, por exemplo, “1303569” no campo de município de ocorrência, enquanto o de residência “1300000”.

Mesmo com regras que tentassem abranger todas as correspondências, observando manualmente as bases, percebeu-se casos custosos e impraticáveis. Como o caso de uma indígena que os campos se mostravam correspondentes, exceto pelo nome; em uma base “SWARI” e em outra, “UARE”. Dessa maneira, mesmo com o pareamento probabilístico, o score foi abaixo do critério estabelecido. E, assim como o caso, outras pessoas tiveram suas informações inutilizadas para a análise de mortalidade infantil no Brasil.

Conforme a tabela 1, foram consideradas as seguintes variáveis com as respectivas bases.

Tabela 1 - Variáveis de Pareamento

Sistemas de Informação	ID	Variáveis para Pareamento	ID	Variáveis para Revisão Manual
SIM	a	Nome	1	Idade da mãe
	b	Sobrenome	2	Sexo
	c	Data de nascimento	3	Município de Ocorrência
	d	Município de Residência	4	Nome da mãe
Sinasc	a	Nome	1	Idade da mãe
	b	Sobrenome	2	Sexo
	c	Data de nascimento	3	Município de Ocorrência
	d	Município de Residência	4	Nome da mãe

A base liberada pela CGIAE/MS apresentou 35.293 registros com 59 variáveis. Para informações que constam tanto da Declaração de Nascido Vivo (DN), como na Declaração de Óbito (DO), considerou-se prioritariamente as informações da DO. Casos com número da DN ou DO inválida foram retirados da análise. Assim, a base para tratamento registrou 26.738 casos com 36 informações, conforme Tabela 2.

Tabela 2 - Descrição de Variáveis

Variável	Descrição	Valores válidos
NUMERODO	Número da declaração de óbito	Identificador
SEXO	Sexo do recém-nascido	1-Masculino; 2-feminino
GRAVIDEZ	Tipo de gravidez	1-única; 2-dupla;3-tripla ou mais; 9-ignorado
IDADEMAE	Idade da mãe	Idade da mãe em anos
TPPOS	Óbito investigado	1-Sim;2-Não
FONTEINV	Fonte de investigação	1- Comitê de Morte Materna e/ou Infantil; 2- Visita domiciliar; 3-Estab Saúde;4- Outros bancos de dados;5-SVO; 6- IML;7- Outra fonte;8- Múltiplas fontes;9- Ignorado
LOCOCOR	Local de ocorrência do óbito	9- Ignorado;1- Hospital;2- Outro estab. saúde;3- Domicílio;4- Via Pública;5- Outros
ATESTANTE	Indica se o médico que assina atendeu o paciente	1- Sim;2- Substituto;3- IML;4- SVO;5- Outros
ESCMAE	Escolaridade da mãe	1 - Fundamental I; 2 - Fundamental II; 3 - Médio; 4 - Superior incompleto; 5 - Superior completo; 9 - Ignorado
PESO	Peso ao nascer em gramas	
APGAR1	Apgar no primeiro minuto de vida	0 a 10
APGAR5	Apgar no 5 min de vida	0 a 10
IDANOMAL	Anomalia identificada	1- Sim; 2-não; 9-ignorado
LOCNASC	Local de nascimento	1 – Hospital; 2 – Outros estabelecimentos de saúde; 3 – Domicílio; 4 – Outros; 5- Aldeia Indígena.
ESTCIVMAE	Estado civil da mãe	1- Solteiro; 2 - Casado; 3 - Viúvo; 4 - Separado judicialmente/divorciado; 5 - Uniao estável; 9-ignorado
RACACORMAE	Raça/cor da mãe	1– Branca; 2– Preta; 3– Amarela; 4– Parda; 5– Indígena
QTDGESTANT	Número de gestações anteriores	
QTDPARTNOR	Qtd. De partos vaginais	
QTDPARTCES	Qtd. De partos cesáreos	
QTDFILVIVO	Número de filhos vivos	
QTDFILMORT	Número de perdas fetais e abortos	
GESTACAO	Semanas de gestação	1– Menos de 22 semanas; 2– 22 a 27 semanas; 3– 28 a 31 semanas; 4– 32 a 36 semanas; 5– 37 a 41 semanas; 6– 42 semanas e mais; 9– Ignorado.
CONSULTAS	Número de consultas de pré-natal	1-Nenhuma;2-de 1 a 3; 3-De 4 a 6; 4-7 e mais; 9- Ignorado
CONSPRENAT	Número de consutas pré-natal	
MESPRENAT	Mês que se iniciou o pré-natal	1– Cefálico; 2– Pélvica ou podálica; 3–Transversa; 9– Ignorado.
TPAPRESENT	Tipo de apresentação do RN	
STTRABPART	Trabalho de parto induzido?	1– Sim; 2– Não; 3– Não se aplica; 9– Ignorado

PARTO	Tipo de parto	1-Vaginal; 2-cesáreo
TPNASCASSI	Nascimento foi assistido por?	1-Médico;2-Enfermagem; 3-parteira;4-Outros; 9-Ignorado
TPROBSON	Código do grupo de Robson	
PARIDADE	Define se é a primeira gravidez ou se teve mais de uma	1 – Multípara; 0- Nulípara
KOTELCHUCK	Avaliação da assistência de pré-natal	
CAUSABAS	Causa básica, Classificação Internacional de Doença (CID), 10a. Revisão	
IDADE	Grupo etário da criança	0-Neonatal precoce; 1-neonatal tardio; 2-Pos-neonatal
RACACOR	Raça/cor da criança	1– Branca; 2– Preta; 3– Amarela; 4– Parda; 5– Indígena
IND_RC	Raça/cor de interesse	0-Não-indígena;1-indígena

Utilizou-se o software estatístico R (4.2.2) para análises exploratória e análise de componentes principais. O software Python (3.10) foi utilizado para modelagem.

3. Processamento/Tratamento de Dados

O relacionamento entre as bases de dados coletadas para a formação do *dataset* final, bem como as informações acerca dos campos foram descritas previamente.

O processo de limpeza está relacionado à qualidade dos dados. Consiste em tratar dados ausentes, inconsistentes, irrelevantes, duplicados ou redundantes. Por outro lado, o de enriquecimento é para criar campos que não constavam na base de dados inicial. A codificação é alteração no formato dos dados para aplicação correta do algoritmo. Técnicas para escolher campos mais significativos, correlação entre variáveis é a operação de *feature engineering*.

Com o objetivo de aplicar algoritmos de ML, são necessários alguns procedimentos listados a seguir. No capítulo “Apêndices” há uma tabela para bases utilizadas e os respectivos códigos.

Caso houvesse informações similares na DN, como na DO, optou-se pela preferência dos registros da DO, para analisar fidedignamente as características do óbito.

Foi necessário excluir números da DN que fossem inválidos, mais ou menos de oito caracteres, campos vazios ou números negativos.

```
#### Retirar numero de DN invalido ####
dados$ncharDN <- nchar(dados$NUMERODN)
dados2 <- dados[!(dados$ncharDN != 8 | is.na(dados$ncharDN) | dados$NUMERODN < 0),]
```

O grupo etário foi dividido entre os grupos neonatal precoce (de 0 a 6 dias de vida), neonatal tardia (de 7 a 27 dias de vida) e o pós-neonatal (acima de 28 dias até um ano de vida).

```
### Organizar idade ####
df$grupo_idade <- ""
df$grupo_idade[which(df$IDADE <=227)] <- "Neonatal Tardia"
df$grupo_idade[which(df$IDADE <=206)] <- "Neonatal Precoce"
df$grupo_idade[which(df$IDADE >= 228)] <- "Pos-neonatal"

df$faixaetaria <- ""
df$faixaetaria[which(df$grupo_idade == "Neonatal Tardia")] <- 1
df$faixaetaria[which(df$grupo_idade == "Neonatal Precoce")] <- 0
df$faixaetaria[which(df$grupo_idade == "Pos-neonatal")] <- 2
```

Os registros com valores nulos foram excluídos. O objetivo do trabalho é traçar agrupamentos com base nas características de cada óbito. A ausência dessas informações limita o mapeamento.

```
## CHECAGEM DE VALORES NULOS POR COLUNA
colSums(is.na(dataset))/length(dataset$NUMERODO)*100
```

A cor/raça foi alterada para dois grupos: indígenas e não-indígenas. Caso não houvesse registro de cor/raça na DO e houvesse declaração na DN, a informação era utilizada afim de evitar respostas nulas.

```
### Separar o DF entre racas indigenas e nao-indigenas ####
df$corraca <- ifelse(is.na(df$RACACOR.x), df$RACACOR.y, df$RACACOR.x)
df$rc_ind <- "" #1 indigena em branco, excluido (tratar antes se nulo)
df$rc_ind[which(df$corraca == 5)] <- 1
df$rc_ind[which(df$corraca != 5 | is.na(df$corraca))] <- 0
```

Campos formais ou redundantes foram excluídos a priori.

```
#### Retirar colunas ####
excluir <- c("NOMEMAE.x", "DTNASC", "SEXO.y", "GRAVIDEZ.y", "IDADEMAE.y", "CODINST.x",  "CODESTAB.x",
             "LINHAA", "LINHAB", "LINHAC", "LINHAD", "LINHAII", "ESCMAE.y", "ESCMAE2010", "CODINST.y", "CODANOMA
             "CODESTAB.y", "CODMUNNASC", "CODMUNOCOR",  "CODMUNRES", "CAUSABAS_O", "DTOBITO")
dados <- dataset[,!(names(dataset)%in% excluir)]
```

A base para a aplicação de componentes principais registrou 19.506 com 32 variáveis.

No software R, utilizando os pacotes "corr", "FactorMineR" e "factoextra", aplicou-se a análise de componentes principais (PCA). Quantidade de gestações, quantidade de filho vivo e morto, peso ao nascer, TPRobson são as variáveis que mais contribuem para variabilidade entre duas dimensões.

4. Análise e Exploração dos Dados

O Brasil registrou 23.262 óbitos infantis. Cerca de 55,7% é de neonatal precoce, seguido de pós-neonatal com 25,4% do casos e 18,8% para neonatal tardio.

As características maternas, apresentadas na tabela 3, contextualizam a situação maternas dos óbitos infantis. Pouco mais de 50% das mães declarou ser solteira, em contrapartida, casadas ou em união estável são 47,2%.

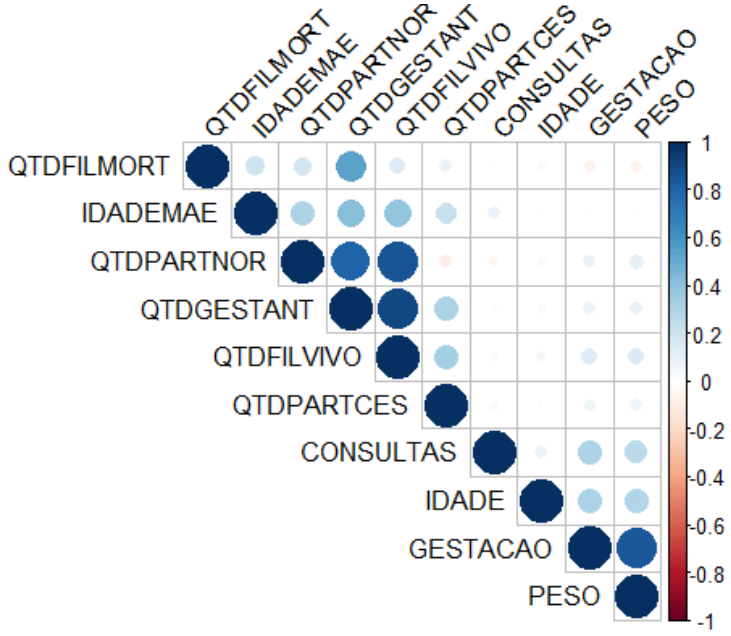
Quanto a escolaridade materna, aproximadamente 55% tiveram acesso ao ensino superior. No que se refere a raça/cor, mais de 60% se declarou como parda ou preta.

Tabela 3 - Características maternas

Variável	N	(%)
Raça/cor mãe		
Parda	12.972	57,80%
Branca	6.876	30,64%
Preta	1.712	7,63%
Indígena	316	1,41%
Amarela	80	0,36%
Não informado	486	2,17%
Estado Civil mãe		
Solteiro	11.285	50,29%
Casado	6.074	27,07%
União estável	4.531	20,19%
Separado judicialmente/divorciado	314	1,40%
Viúvo	46	0,20%
Não informado	192	0,86%
Escolaridade da mãe		
Superior Incompleto	12.334	54,96%
Médio	4.209	18,76%
Superior Completo	3.543	15,79%
Fundamental II	803	3,58%
Fundamental I	520	2,32%
Não informado	1.033	4,60%

Dentre as variáveis numéricas, observou-se a correlação entre elas (Figura 1). Os campos de quantidade de filhos vivos e a quantidade de parto normal apresentou correlação positiva, assim como as variáveis de número de gestação e o peso do recém-nascido.

Figura 1 - Correlação entre variáveis



Devido ao número de variáveis, decidiu-se por aplicar a análise de componentes principais a fim de detectar características que mais contribuem para a variabilidade dos óbitos infantis.

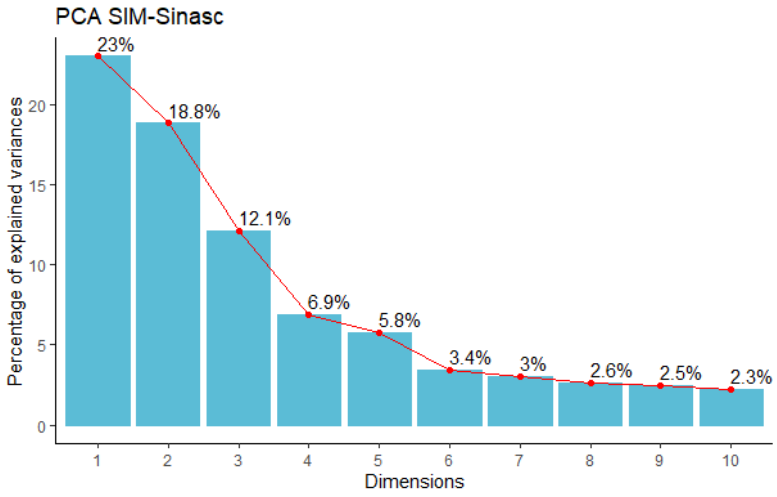
A análise de componentes principais é uma técnica da estatística multivariada que consiste em transformar um conjunto de variáveis originais em outro conjunto de variáveis de mesma dimensão denominadas de componentes principais [6]. A análise de componentes principais é associada à redução de massa de dados, com menor perda possível da informação.

```
> ## COMPONENTES
> data.pca <- princomp(corr_matrix)
> summary(data.pca)
Importance of components:
                          Comp.1    Comp.2    Comp.3     Comp.4     Comp.5     Comp.6    Comp.7     Comp.8     Comp.9    Comp.10
Standard deviation     0.5876582 0.5315959 0.4268116 0.32128685 0.29444473 0.22635867 0.2128838 0.19830976 0.19381925 0.18392845
Proportion of Variance 0.2301099 0.1882993 0.1213831 0.06878148 0.05776878 0.03414129 0.0301975 0.02620437 0.02503107 0.02254153
Cumulative Proportion  0.2301099 0.4184092 0.5397923 0.60857377 0.66634255 0.70048384 0.7306813 0.75688571 0.78191678 0.80445831
```

Cada componente explica uma porcentagem da variância total no conjunto de dados. Na seção Proporção Cumulativa, o primeiro componente principal explica 23% da variância total. A segunda explica 41,8% da variância total.

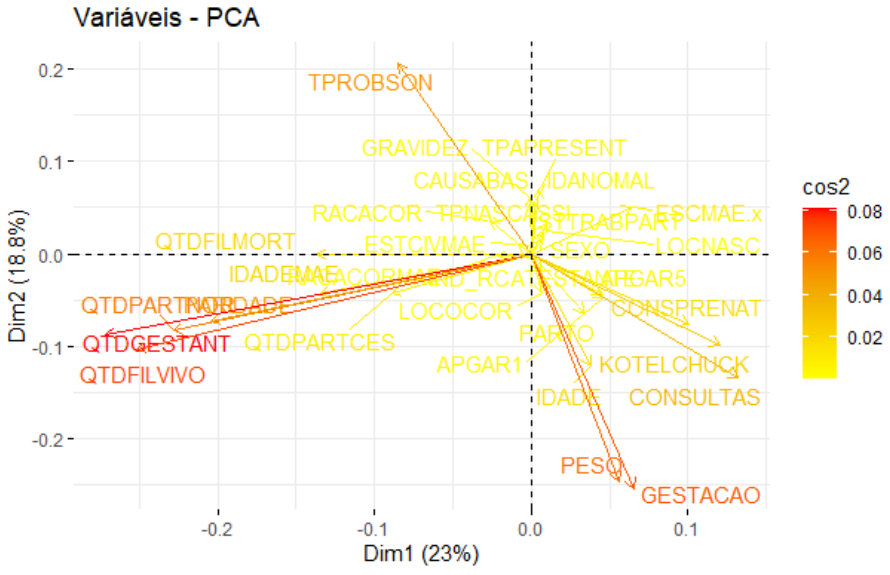
A figura 2 expõe a importância de cada componente principal e pode ser usado para determinar o número de componentes principais a serem retidos. Este gráfico mostra os autovalores em uma curva descendente, do maior para o menor. Os dois primeiros componentes podem ser considerados os mais significativos.

Figura 2 - Scree Plot PCA



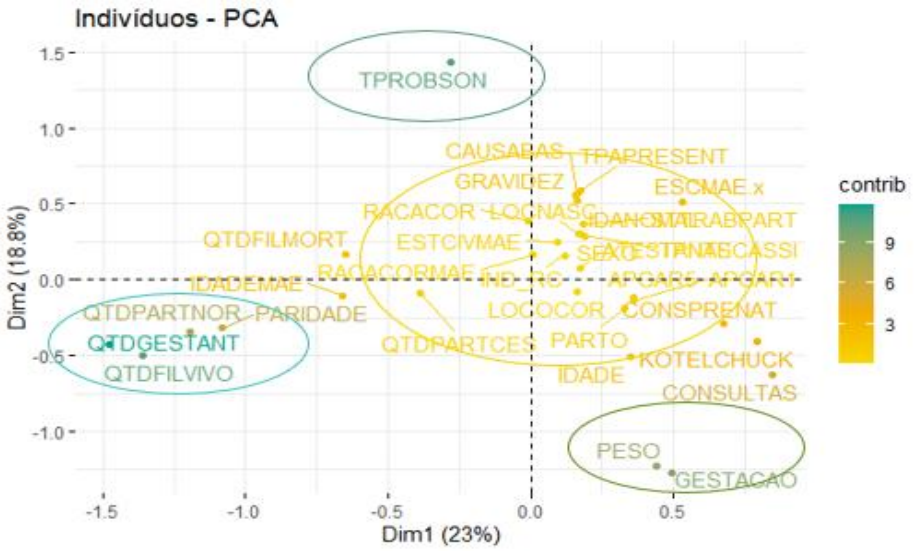
É possível visualizar as semelhanças e diferenças entre as amostras, além de mostrar o impacto de cada atributo em cada um dos componentes principais (Figura 3). Todas as variáveis que são agrupadas estão positivamente correlacionadas entre si, e é o caso, por exemplo, de "Peso" e "Gestacao". As variáveis negativamente correlacionadas são exibidas nos lados opostos da origem do biplot, como "TPROBSON" e "QTDFILVIVO". Cos2 é chamado de cosseno quadrado (coordenadas quadradas) e corresponde à qualidade de representação das variáveis. Valores altos de Cos2 significam uma boa representação da variável naquele componente.

Figura 3 - Contribuição variáveis



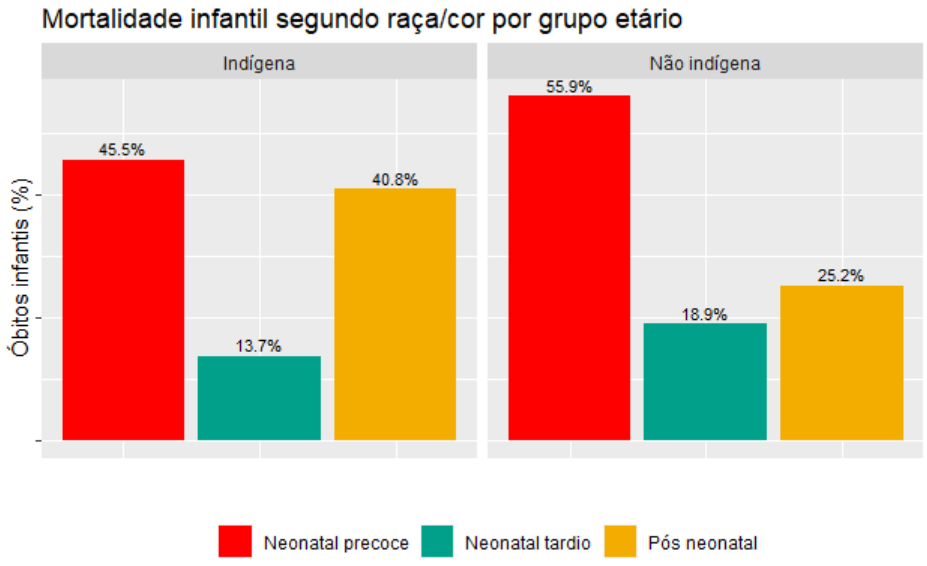
Quanto aos indivíduos, a análise será focada no cos2 e nas contribuições dos indivíduos para os dois primeiros componentes principais (PC1 e PC2). Uma boa representação dos indivíduos no componente principal, se estão posicionados próximos à circunferência do círculo de correlação.

Figura 4 - Contribuição indivíduos



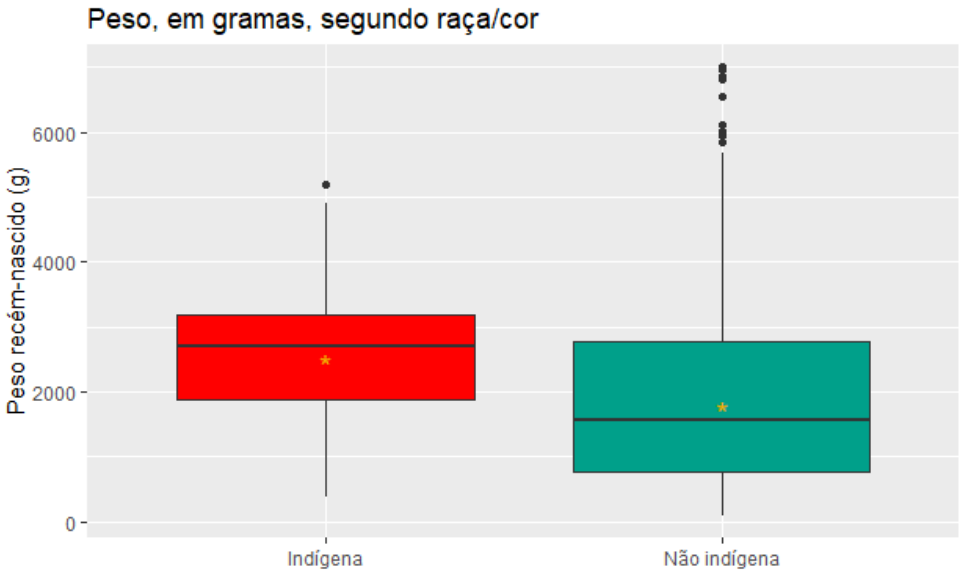
A mortalidade infantil segundo raça/cor por grupo etário indica que há um padrão distinto entre indígenas e não-indígenas (Figura 5). A frequência maior nos dois grupos é de neonatal precoce, seguido do pós neonatal. Enquanto para indígenas é cerca de 40,8%, para não indígenas a proporção é de 25,2%.

Figura 5



O peso do recém-nascido indígena é maior que o peso de recém-nascidos não-indígenas. A própria média dos dois grupos segue esse comportamento.

Figura 6



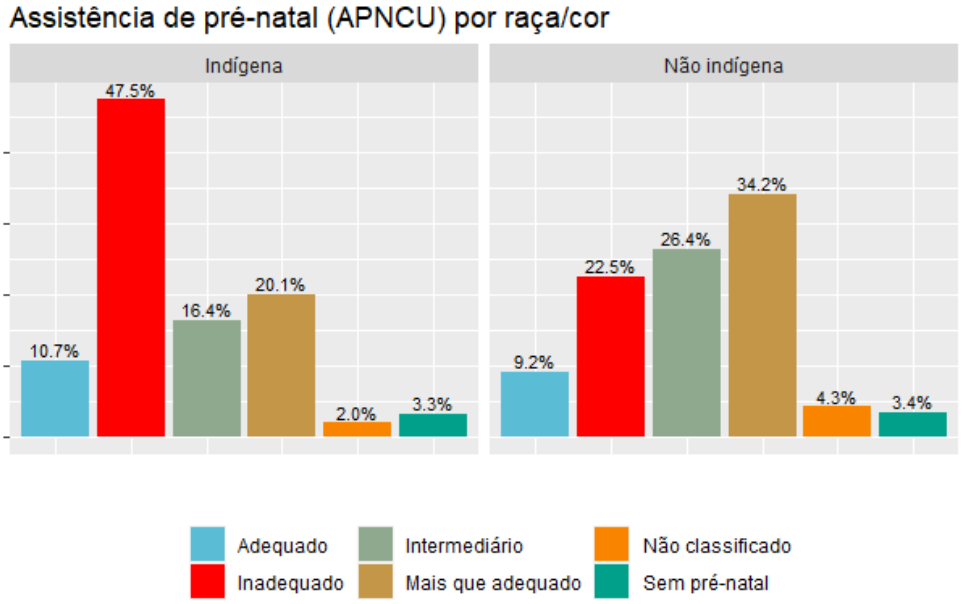
O Índice de Kotelchuck (IK) [7] considera “adequado” o pré-natal iniciado antes da 13ª semana de gestação e se houver uma ou mais consultas para gestação de 13 ou menos semanas; duas ou mais consultas para gestação de 14 a 17 semanas; três ou mais consultas para a gestação de 18 a 21 semanas; quatro ou mais consultas para gestação de 22 a 25 semanas; cinco ou mais consultas para gestação de 26 a 29 semanas; seis ou mais consultas para gestação de 30 a 31 semanas; sete ou mais consultas para gestação de 32 a 33 semanas; oito ou mãos consultas para gestação de 34 a 35 semanas e nove ou mais consultas para gestação de 36 ou mais semanas.

É considerado “inadequado” se não houver consultas em gestação de 14 a 21 semanas; uma consulta ou menos em gestação de 22 a 29 semanas; duas ou menos consultas em gestação de 30 a 31 semanas; três ou menos consultas em gestação de 32 a 33 semanas; e quatro ou menos consultas em gestação a partir de 34 semanas. O cuidado é classificado como “intermediário” para todas as outras combinações que se enquadram nas especificidades anteriores. Apesar de importante para avaliação da utilização dos serviços de pré-natal, esse índice não permite que se avaliem o conteúdo e a qualidade da assistência.

Comparando os dois grupos, percebe-se a alta frequência de pré-natal inadequado para indígenas e mais que adequado para não indígenas. O

acompanhamento do pré-natal e da atenção ao parto é reconhecido atualmente como importante estratégia para prevenir ou reduzir o risco de mortalidade, tanto para a gestante como para a criança.

Figura 7



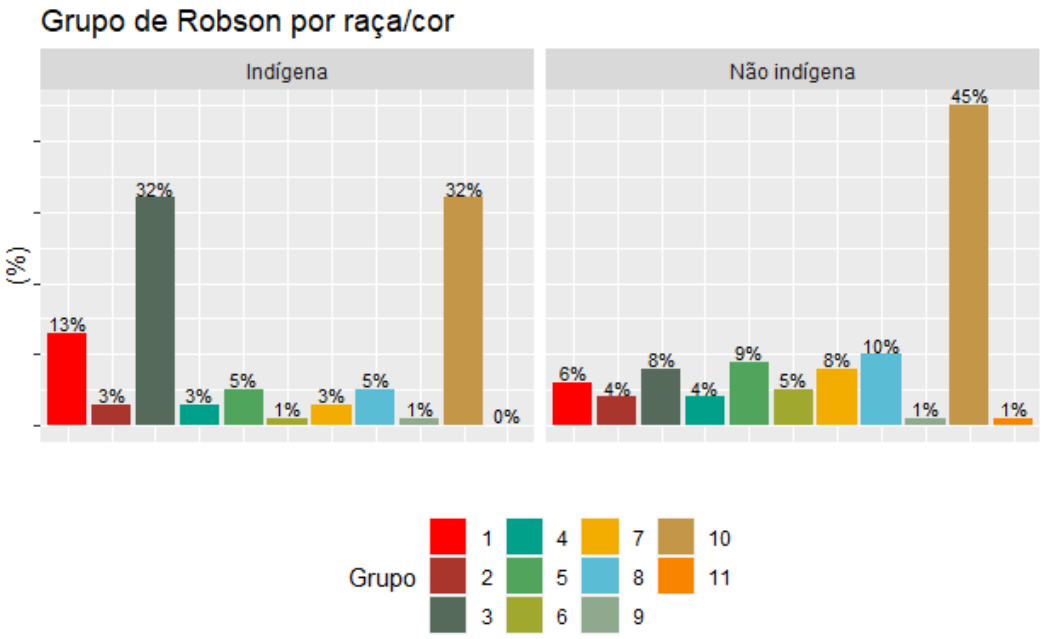
A classificação de dez grupos de Robson [8] baseia-se em parâmetros obstétricos simples, como paridade, cesárea anterior, idade gestacional, indução do parto, relação temporal entre realização de cesárea e o início do trabalho de parto, apresentação fetal e número de fetos, combinadas de acordo com o exposto na tabela 4, não envolvendo a indicação obstétrica de cesárea.

Tabela 4 - Grupos de Robson

Grupo 1	Nulípara, gestação única, cefálica, >36semanas, em trabalho de parto espontâneo
Grupo 2	Nulípara, gestação única, cefálica, >36semanas, com indução ou cesárea anterior ao trabalho de parto
Grupo 3	Multípara (sem antecedente de cesárea), gestação única, cefálica, >36 semanas, em trabalho de parto espontâneo
Grupo 4	Multípara (sem antecedente de cesárea), gestação única, cefálica, >36 semanas, com indução ou cesárea realizada antes do início do trabalho de parto
Grupo 5	Com antecedente de cesárea, gestação única, cefálica, >36 semanas
Grupo 6	Todos os partos pélvicos em nulíparas
Grupo 7	Todos os partos pélvicos sem multíparas
Grupo 8	Todas as gestações multíparas
Grupo 9	Todas as apresentações anormais
Grupo 10	Todas as gestações únicas, cefálicas, >36 semanas

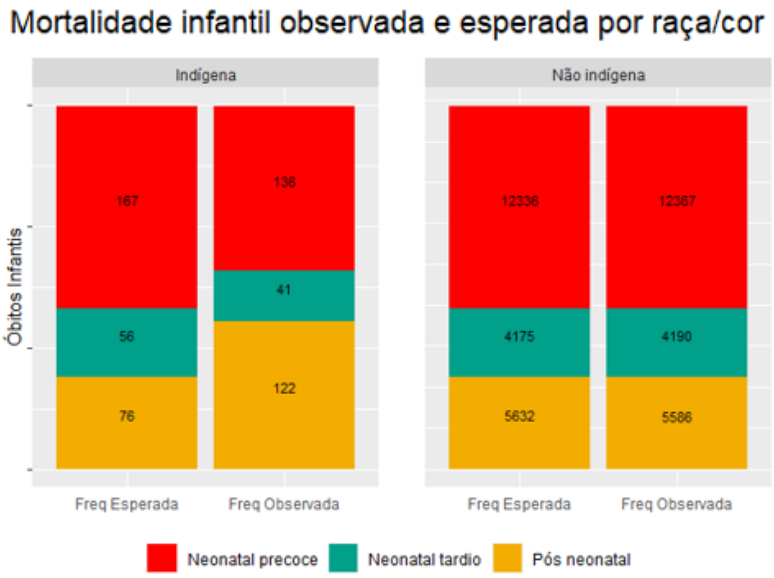
Para os indígenas, 32% estão no grupo 3 e 32% no grupo 10. Para os não-indígenas, 45% estão no grupo 10, seguido do grupo 8 com 10%.

Figura 8



Como há indícios de uma diferença significativa no perfil entre indígenas e não indígenas, realizou-se o Teste do Qui-Quadrado. A hipótese nula é de que há independência entre os grupos de raça/cor e o grupo etário infantil. O X^2 é de 38,07 com p-valor menor que 0,01. Comparando as frequências esperadas e observadas de acordo com a raça/cor, o grupo indígena tem diferença significativa entre os valores (Figura 9).

Figura 9



5. Modelos de Machine Learning

Conforme tratado, criaremos modelos para agrupamento do perfil de óbitos infantis entre indígenas e não indígenas em 2019 no Brasil.

Após a verificação da consistência dos dados, a base de dados deve ser convertida para um formato que possa ser utilizado pela aplicação onde será realizada a mineração. Neste momento são aplicados um ou mais métodos específicos para extração de regras ou padrões.

O aprendizado não-supervisionado ocorre quando a variável alvo não é conhecida na base de dados e a mineração categoriza apenas similaridade entre os dados. Os algoritmos de clusterização podem ser divididos em duas categorias: partitivos ou hierárquicos. Os algoritmos partitivos dividem o conjunto de dados em k clusters e produzem agrupamentos simples, tentando fazer os clusters tão compactos e separados quanto possível. Entretanto, quando existem grandes diferenças nos tamanhos e geometrias dos diferentes clusters, podem dividir desnecessariamente grandes clusters para minimizar a distância total calculada.

A análise por Silhouette mede o quão bem um ponto se encaixa em um cluster. Neste método um gráfico é feito medindo quão perto os pontos de um cluster estão dos pontos de outro cluster mais próximo. O coeficiente de Silhouette quando próximo de +1, indica que os pontos estão muito longe dos pontos do outro cluster, e quando próximo de 0, indica que os pontos então muito perto ou até interseccionando um outro cluster.

Diferentes métricas de desempenho são usadas para avaliar diferentes algoritmos de aprendizado de máquina. Tanto o critério de informação de Akaike (AIC) quanto o critério de informação bayesiano (BIC) são sistemas de pontuação para comparações de modelos em estatísticas clássicas que lidam com modelos com diferentes números de parâmetros livres. O índice de Dunn é identificar conjuntos de clusters que são compactos, com uma pequena variação entre os membros do cluster, e bem separados, onde as médias dos diferentes clusters estão suficientemente distantes, em comparação com o dentro da variância do cluster. Quanto maior o valor do índice Dunn, melhor é o agrupamento.

5.1 Density-Based Spatial Clustering Of Application with Noise (DBSCAN)

DBSCAN é um modelo de cluster baseado em densidade e os modelos procuram os dados espaço para áreas de diferentes densidades de pontos de

dados. Eles separam diferentes regiões de densidade e atribui pontos de dados dentro dessas regiões no mesmo cluster. Os dois principais benefícios do modelo DBSCAN são que não precisa informar o número de clusters inicial e é robusto contra outliers. DBSCAN requer apenas dois parâmetros: Epsilon e minPoints.

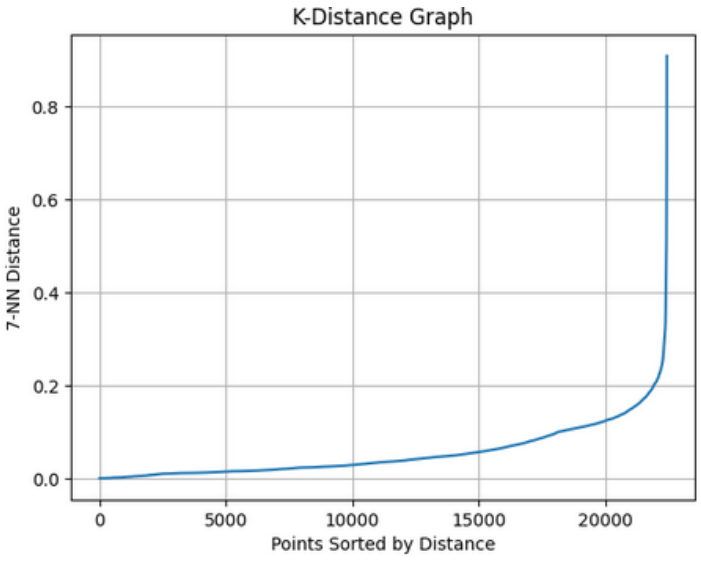
Normalizando os valores

```
ms = MinMaxScaler()
cols = df.columns

X = ms.fit_transform(df)
X = pd.DataFrame(X, columns = [cols])
```

```
db = DBSCAN(eps = 0.4, min_samples = 7).fit(X)
core_samples_mask = np.zeros_like(db.labels_, dtype=bool)
core_samples_mask[db.core_sample_indices_] = True
labels = db.labels_

# Number of clusters in labels, ignoring noise if present

n_clusters_ = len(set(labels)) - (1 if -1 in labels else 0)
n_noise_ = list(labels).count(-1)
print('Número de Clusters : ', n_clusters_)
print('Número de Outliers : ', n_noise_)
```

```
Número de Clusters :  25
Número de Outliers :  51
```

EPS	Samples	Silhouette Score
0,20	7	-0,45
0,20	8	-0,43
0,30	7	-0,40
0,30	8	-0,39
0,40	7	-0,33
0,40	8	-0,34



5.2 Gaussian Mixture Model (GMM)

O Gaussian Mixture Model (GMM) é um modelo probabilístico que assume que todos os pontos de dados são gerados a partir de uma mistura de distribuições gaussianas com parâmetros desconhecidos. Um modelo de mistura gaussiana pode ser usado para clustering, que é a tarefa de agrupar um conjunto de pontos de

dados em clusters. GMMs podem ser usados para encontrar clusters em conjuntos de dados onde os clusters podem não estar claramente definidos.

Normalizando os valores

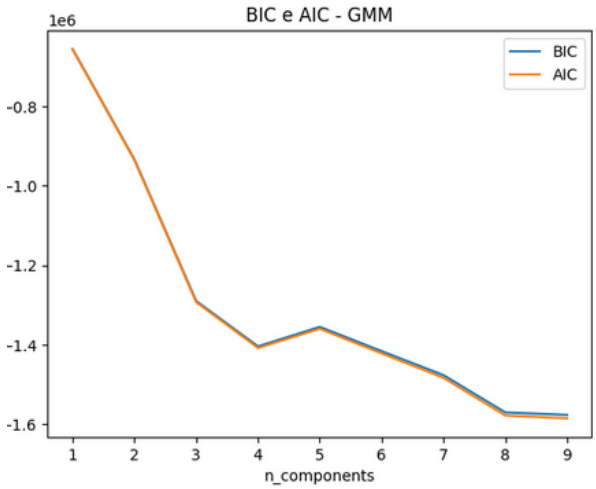
```
ms = MinMaxScaler()
cols = df.columns

X = ms.fit_transform(df)
X = pd.DataFrame(X, columns = [cols])
```

```
range_n_components = [2, 3, 4, 5, 6, 7, 8]

for num_componentes in range_n_components:

    Gmm = GaussianMixture(n_components=num_componentes).fit(X)

    silhouette_avg = np.exp(Gmm.score_samples(X)).sum()
    print("Para {0} grupos, a silhouette score é de {1}".format(num_componentes, silhouette_avg))
```

Feature Scaling

```
n_components = np.arange(1, 10)
models = [GaussianMixture(n, covariance_type='full', random_state=0).fit(X)
          for n in n_components]

plt.plot(n_components, [m.bic(X) for m in models], label='BIC')
plt.plot(n_components, [m.aic(X) for m in models], label='AIC')
plt.legend(loc='best')
plt.title('BIC e AIC - GMM')
plt.xlabel('n_components');
```

Número Componentes	GMM Score
2	4,53
3	4,53
4	1,87
5	1,86
6	7,04
7	2,65
8	6,87



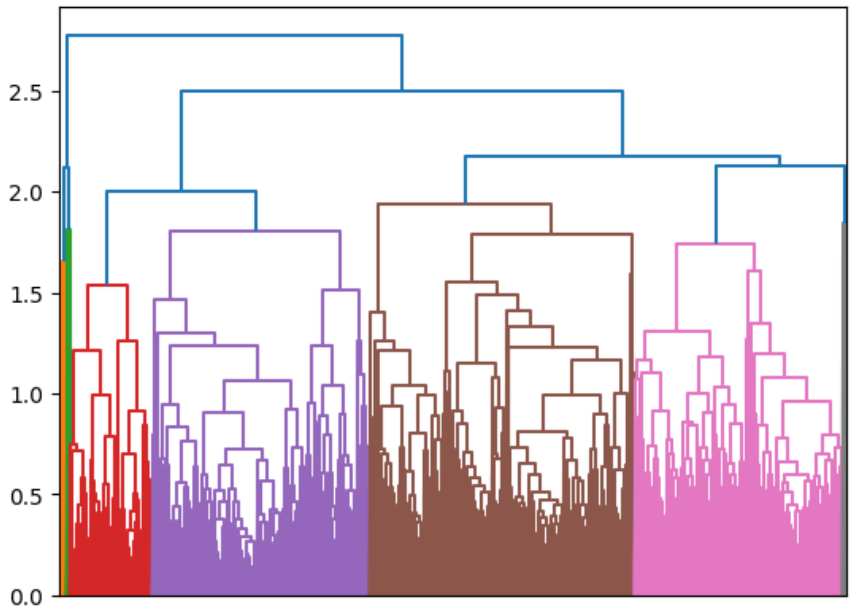
5.3 Hierárquico

Os agrupamentos hierárquicos são formados por nós, que posteriormente são mesclados de forma repetitiva até serem fundidos em um único nó por meio de uma função matemática de similaridade ou distância escolhida, conhecida como função de ligação, como a distância euclidiana, por exemplo. Segundo Bussab et al. (1990) [9], no método da ligação completa, a dissimilaridade entre dois grupos é definida como aquela apresentada pelas parcelas de cada grupo que mais se parecem.

```
ms = MinMaxScaler()
cols = df.columns

X = ms.fit_transform(df)
X = pd.DataFrame(X, columns = [cols])
```

```
linkage_data = linkage(X, 'complete')
dendrogram(linkage_data)
plt.figure(figsize=(10, 7))
```

Número de Clus-ters	Silhouette Score
2	0,37
3	0,36
4	0,33
5	0,32
6	0,32
7	0,36



5.4 K-Means

É um algoritmo usado quando se tem dados não rotulados, que são dados sem categorias ou grupos definidos. O algoritmo segue uma maneira fácil ou simples de classificar um determinado conjunto de dados por meio de um determinado número de clusters, fixado a priori. O algoritmo K-Means funciona iterativamente para atribuir cada ponto de dados a um dos grupos K com base nos recursos fornecidos. Os pontos de dados são agrupados com base na similaridade de recursos. Quanto menor o valor da *inertia*, melhor o ajuste do modelo. Por outro lado, quanto maior o Índice de Davies, melhor o ajuste.

No método do cotovelo, é preciso rodar seu algoritmo de *clustering* com alguns valores, por exemplo, de 1 a 10. Calcular a função de custo, a soma dos quadrados das distâncias internas dos clusters, e traçá-la em um gráfico. O melhor número para a quantidade de clusters é quando a adição de um novo cluster não muda significativamente a função de custo. Isso geralmente acontece no "cotovelo" da linha.

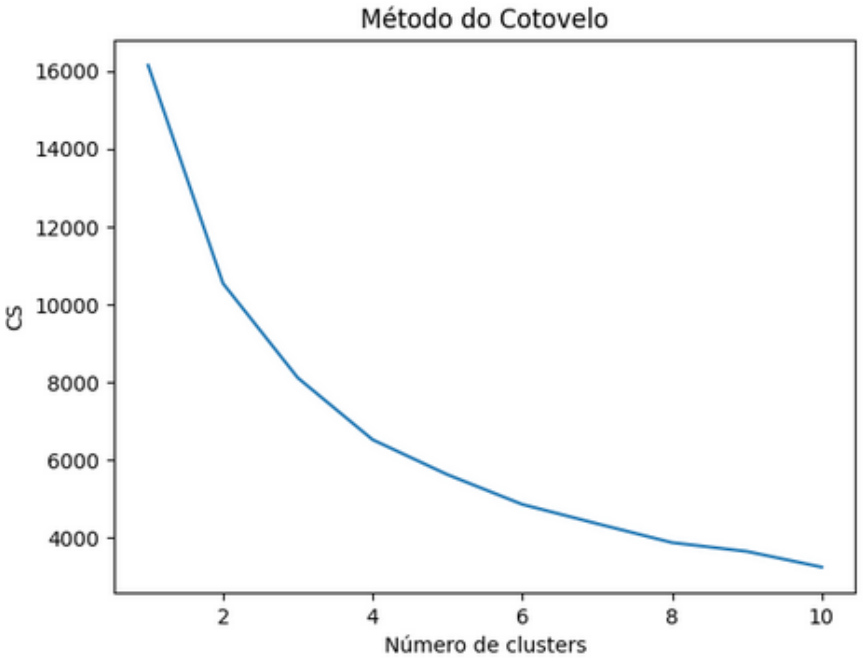
```
kmeans = KMeans(n_clusters=3,random_state=10)

kmeans.fit(X)

labels = kmeans.labels_

X['Cluster_Id'] = kmeans.labels_
print(davies_bouldin_score(X, labels))
```

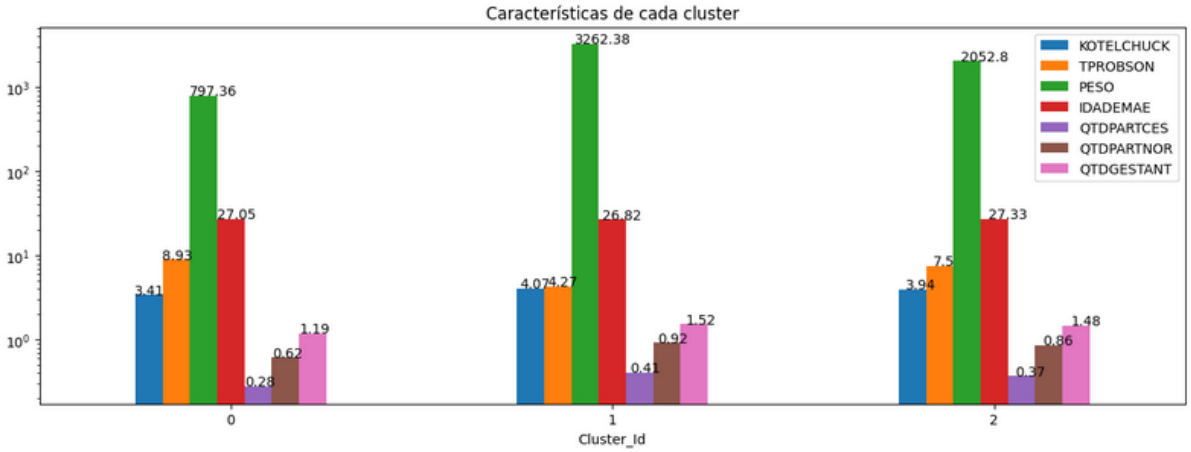
```
0.40100145997161496
```

```
kmeans.inertia_
```

```
8113.7849070413085
```

Número Clusters	Silhouette Score
2	0,68
3	0,62
4	0,58
5	0,56
6	0,54
7	0,53
8	0,53

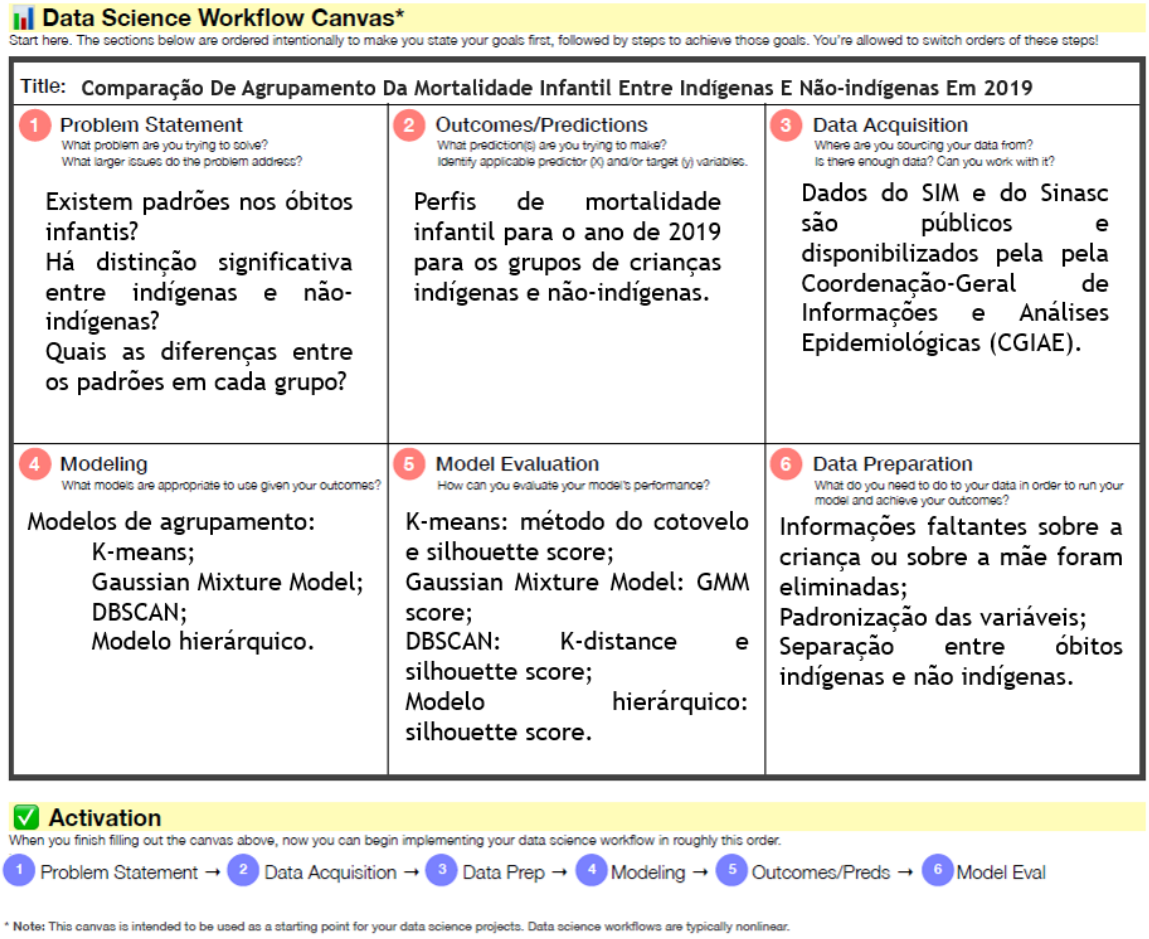

Método do Cotovelo
CS
16000
14000
12000
10000
8000
6000
4000
2
4
6
8
10
Número de clusters


Características de cada cluster
KOTELCHUCK
TPROBSON
PESO
IDADEMAE
QTDPARTCES
QTDPARTNOR
QTDGESTANT
3.41
8.93
797.36
27.05
0.28
0.62
1.19
4.07
4.27
3262.38
26.82
0.41
0.92
1.52
3.94
7.5
2052.8
27.33
0.37
0.86
1.48
0
1
2
Cluster_Id

7. Resultados

Uma boa forma de sinterizar o que foi feito neste trabalho é utilizar o modelo de fluxo de trabalho de ciência de dados proposto por Vasandani (2019), conforme Figura 10.

Figura 10 - Workflow



Data Science Workflow Canvas*

Start here. The sections below are ordered intentionally to make you state your goals first, followed by steps to achieve those goals. You're allowed to switch orders of these steps!

Title: Comparação De Agrupamento Da Mortalidade Infantil Entre Indígenas E Não-indígenas Em 2019

1 Problem Statement What problem are you trying to solve? What larger issues do the problem address?	2 Outcomes/Predictions What prediction(s) are you trying to make? Identify applicable predictor (X) and/or target (y) variables.	3 Data Acquisition Where are you sourcing your data from? Is there enough data? Can you work with it?
Existem padrões nos óbitos infantis? Há distinção significativa entre indígenas e não-indígenas? Quais as diferenças entre os padrões em cada grupo?	Perfis de mortalidade infantil para o ano de 2019 para os grupos de crianças indígenas e não-indígenas.	Dados do SIM e do Sinasc são públicos e disponibilizados pela pela Coordenação-Geral de Informações e Análises Epidemiológicas (CGIAE).
4 Modeling What models are appropriate to use given your outcomes?	5 Model Evaluation How can you evaluate your model's performance?	6 Data Preparation What do you need to do to your data in order to run your model and achieve your outcomes?
Modelos de agrupamento: K-means; Gaussian Mixture Model; DBSCAN; Modelo hierárquico.	K-means: método do cotovelo e silhouette score; Gaussian Mixture Model: GMM score; DBSCAN: K-distance e silhouette score; Modelo hierárquico: silhouette score.	Informações faltantes sobre a criança ou sobre a mãe foram eliminadas; Padronização das variáveis; Separação entre óbitos indígenas e não indígenas.

Activation

When you finish filling out the canvas above, now you can begin implementing your data science workflow in roughly this order.

1 Problem Statement → 2 Data Acquisition → 3 Data Prep → 4 Modeling → 5 Outcomes/Preds → 6 Model Eval

* Note: This canvas is intended to be used as a starting point for your data science projects. Data science workflows are typically nonlinear.

Conceptualized by Jasmine Vasandani using notes from General Assembly's Data Science Immersive. Format inspired by Business Model Canvas.

O algoritmo DBSCAN apresentou o índice de silhueta negativo para os distintos valores de épsilon. O algoritmo apontou 25 clusters para os dados apresentados. Isso inviabiliza qualquer tipo de procedimento para abordar e cuidar dos grupos. Por isso, decidiu-se não seguir com essa abordagem.

O algoritmo Hierárquico exigiu um recurso computacional além do esperado para identificar os grupos apontados. Os valores do coeficiente da silhueta indicam que não propriamente distinção eficiente entre os clusters.

O algoritmo GMM apresentou o mesmo comportamento tanto para AIC quanto no BIC, ambos minimizados com 8 ou 9 clusters.

O algoritmo K-Means apresentou o melhor resultado para o índice da silhueta para 2 e 3 clusters. O agrupamento ainda é factível com o proposto de traçar perfis para acompanhamento do grupos.

Há três grupos para indígenas e outros três para não indígenas (Figura 11). Os atributos constatados em cada grupo de raça/cor sugerem que o contexto de cuidados durante a gestação é um fator determinante. Enquanto para os não-indígenas os atributos acerca do pré-natal se diferenciam entres os três clusters, para os não indígenas não há exatidão nesse quesito.

O peso do recém-nascido é uma característica que diferencia os três clusters de não indígenas, mas não entre os indígenas que vieram a óbito.

Figura 11 - Clusters por raça/cor

O uso da informação capaz de demonstrar ou aproximar as diferentes realidades locais permitem melhorar cada vez os sistemas de saúde envolvidos no diagnóstico, na notificação e na coleta dos dados sobre mortalidade e, em especial, a mortalidade infantil. A inclusão de variáveis e otimização do preenchimento da DO pode aumentar a confiabilidade, representatividade e melhor aderência aos modelos.

Estudos avaliativos com foco especial nos processos de trabalho podem contribuir para identificar possíveis fatores associados à morte infantil evitável, tais

como a qualidade técnica da atenção prestada, a oportunidade de acesso a procedimentos especializados, entre outros.

Apesar do massivo volume de dados existente, o resultado de acordo com as métricas de avalição são claras que o modelo é apenas razoável.

Considerando a população indígena no Brasil, a deficiente qualidade de dados tem sido um obstáculo para compreender o comportamento das variáveis demográficas, suas características e determinantes.

Em um algoritmo de clusterização, os dados não são divididos tendo em vista apenas um atributo, mas sim com a influência e relação entre estes atributos. Por este motivo, a análise do cluster se mostra extremamente complicada, uma vez que também não se deve observar cada atributo individualmente, mas sim os clusters como um todo e os dados pertencentes a cada um.

Para futuros trabalhos, identificar padrões espaciais, visto que o que determina o agrupamento em diferentes contextos históricos e geográficos.

8. Links

Link para o vídeo: https://youtu.be/vjR5Yg5lMTE

Link para o repositório: https://github.com/naharift/PUCMINAS

REFERÊNCIAS

[1] Bottaci, L., Drew, P. J., Hartley, J. E., Hadfield, M. B., Farouk, R., Lee, P. W., et al. (1997). Artificial Neural Networks Applied to Outcome Prediction for Colorectal Cancer Patients in Separate Institutions. The Lancet 350, 469–472.

[2] World Health Organization, UNICEF, United Nations Population Fund and The World Bank, Trends in Maternal Mortality: 2000 to 2017 WHO, Geneva, 2019.

[3] UNITED NATIONS. The Millennium Development Goals 2015. New York, 2015. Disponível em: Acesso em: novembro. 2021.

[4] Victora CG, Barros FC. Infant mortality due to perinatal causes in Brazil: trends, regional patterns and possible interventions. Sao Paulo Med J. 2001;119:33-42.

[5] Fellegi, I. P., and Sunter, A. B. (1969), "A Theory for Record Linkage," Journal of the American Statistical Association.

[6] JOHNSON, R. A.; WICHERN, D. W. Applied multivariate statistical analysis. 4th ed. Upper Saddle River, New Jersey: Prentice-Hall, 1999, 815 p.

[7] KOGAN, M. D.; MARTIN, J. A.; ALEXANDER, G. R.; KOTELCHUCK, M.; VENTURA, S. J.; FRIGOLETTO, F. D. The Changing Pattern of Prenatal Care Utilization in the United States, 1981-1995, Using Different Prenatal Care Indices. JAMA, v.279, n.20, p.1623-8, 1998

[8] Vogel et al. Use of the Robson classification to assess caesarean section trends in 21 countries: a secondary analysis of two WHO multicountry surveys. Lancet Glob Health 2015; 3: e260–70.

[9] BUSSAB, W. O.; MIAZAKI, E. S.; ANDRADE, D. Introdução à análise de agrupamentos. In: SIMPÓSIO NACIONAL DE PROBABILIDADE E ESTATÍSTICA (SINAPE), 9., 1990, Anais... São Paulo: Associação Brasileira de Estatística, 1990. p.72-75

APÊNDICE

Arquivos e Scripts utilizados

	Processamento	Base de dados
Arquivo gerado pelo Ministério da Saúde	tratamento.R	dados_sim_sinasc_2019.csv
Arquivo PCA	pca.R	sim_sinasc_tratada.csv
Arquivo EDA	eda.R	base_final.csv
Notebook GMM	gmm.ipynb	base_final.csv
Notebook DBSCAN	dbscan.ipynb	base_final.csv
Notebook Kmeans	kmeans.ipynb	base_final.csv
Notebook Hierarquico	hierarchical.ipynb	base_final.csv