**PONTIFÍCIA UNIVERSIDADE CATÓLICA DE MINAS GERAIS  
NÚCLEO DE EDUCAÇÃO A DISTÂNCIA**

**Pós-graduação *Lato Sensu* em Ciência de Dados e Big Data**

**Herberton Candido Souza**

**Modelo de Prevenção à Mortalidade no Parto**

Belo Horizonte

2019

**Herberton Candido Souza**

**MODELO DE PREVANÇÃO À MORTALIDADE NO PARTO**

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

2019**SUMÁRIO**

[1. Introdução 4](#_Toc22444946)

[1.1. Contextualização 4](#_Toc22444947)

[1.2. O problema proposto 4](#_Toc22444948)

[2. Coleta de Dados 6](#_Toc22444949)

[3. Processamento/Tratamento de Dados 7](#_Toc22444950)

[3.1. raw-data 7](#_Toc22444951)

[3.2. stage-data 8](#_Toc22444952)

[3.3. model-data 15](#_Toc22444953)

[4. Análise e Exploração dos Dados 17](#_Toc22444954)

[5. Criação de Modelos de Machine Learning 23](#_Toc22444955)

[5.1. Resultados do Treino do Modelo: 26](#_Toc22444956)

[6. Apresentação dos Resultados 30](#_Toc22444957)

[6.1. Métricas de Qualidade do Modelo 30](#_Toc22444958)

[7. Links 32](#_Toc22444959)

[REFERÊNCIAS 33](#_Toc22444960)

# 1. Introdução

## Contextualização

Este material é o resultado de um estudo que tenta elaborar um questionário básico que serve de entrada para um modelo estatístico e que tenta prever uma possível morte ao final de uma gestação. Este estudo poderá ser utilizado por médicos para prevenir possíveis fatalidades que aconteceriam no momento do parto.

## 1.2. O problema proposto

### Por quê?

Na área da saúde é sempre importante descobrir um eventual problema de maneira antecipada para que haja tempo hábil de contorno evitando que alguma fatalidade ocorra. Descobrir se há chances de uma eventual morte do bebê no momento do seu parto pode ser uma poderosa ferramenta de prevenção e preservação da vida humana.

### Para quem?

Este estudo pode servir para mães, pais, médicos, enfermeiros, estudantes da área da saúde e afins. Todavia, restringe-se à população que mora nos Estados Unidos da América (EUA) devido ao fato dos dados utilizado pelo modelo serem de origem norte-americana.

### O que?

Este estudo trata da apresentação de um modelo estatístico que utilizou como base grandes massas de dados de domínio público para ser criado e que consegue prever uma morte no parto através de entradas simples relacionadas às características dos pais (idade, vícios, raça e afins) e do momento em que vivem (IDH).

### Onde?

O modelo estatístico foi ajustado, ou calibrado, com base em dados públicos dos EUA disponibilizados através da internet. Desse modo, este modelo estatístico serve apenas para gestações de pais que residem em solo norte-americano.

### Quando?

Foram analisados dados de nascimento (partos) de 1969 a 2008 e dados de IDH (Índice de Desenvolvimento Humano) por Estado de 1990 a 2017.

# 2. Coleta de Dados

Abaixo segue o link de uma planilha contendo todas as origens e os metadados dos dados nas seguintes visões: Databases, Sechemas, Table e Columns:

<https://github.com/herberton/tcc.cdbd.puc.mg/raw/master/doc/metadata.xlsx>

# 3. Processamento/Tratamento de Dados

Na análise foi seguido o seguinte processo:

# 3.1. raw-data

As seguintes tabelas foram utilizadas como dado bruto da análise:

* natality

Tabela que contém registros de 137.826.763 partos ocorridos nos e EUA desde o ano de 1969 até o ano 2008 separados por estado.

* fips\_codes\_states
* gdl\_sub\_national\_hdi\_data
* gdl\_educational\_index\_data
* gdl\_expected\_years\_schooling\_data
* gdl\_health\_index\_data
* gdl\_income\_index\_data
* gdl\_life\_expectancy\_data
* gdl\_log\_gross\_national\_income\_per\_capita\_data
* gdl\_mean\_years\_schooling\_data
* gdl\_population\_size\_in\_thousands\_data

## 3.2. stage-data

As seguintes tabelas/views são resultantes da preparação dos dados brutos. Abaixo também descreverei qual é o tratamento dado aos atributos utilizados na análise:

* **prepared\_natality**

View criada a partir da tabela bruta (raw-data) natality.

* + **birth\_year:**

Coluna gerada a partir da coluna year da tabela natality. Pode conter valor nulo, vazio ou que não representa um ano válido.

* + **birth\_month:**

Coluna gerada a partir da coluna month da tabela natality. Pode conter valor nulo, vazio ou que não representa um mês válido.

* + **birth\_day:**

Coluna gerada a partir da coluna day da tabela natality. Pode conter valor nulo, vazio ou que não representa um mês válido.

* + **birth\_date:**

Coluna gerada a partir das colunas year, month e day da tabela natality. Apenas terá valor se essas colunas representarem uma data válida, caso contrário esta coluna é nula.

* + **birth\_fu:**

Coluna gerada a partir da coluna state da tabela natality. Todavia, valores em branco na tabela de origem são considerados como nulos nesta view.

* + **is\_male:**

Coluna gerada a partir da coluna is\_male da tabela natality.

* + **child\_race:**

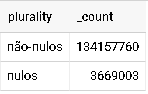
Coluna gerada a partir da coluna child\_race da tabela natality. Quando o valor da origem for nulo, nesta view o valor será o número 9.

* + **weight\_pounds:**

Coluna gerada a partir da coluna weight\_pounds da tabela natality.

* + **plurality:**

Coluna gerada a partir da coluna plurality da tabela natality. Esta coluna representa a quantidade de filhos que estava na barriga da mãe, independente se estes filhos nascerão vivos ou não. Esta coluna é importante para a análise, quando o valor for null na origem é atribuído o valor 1 para esta view, pois entende-se que cada registro da tabela representa uma gravidez de ao menos 1 filho. A quantidade exata de filhos é importante na análise, todavia retirar as colunas que não possuem valor cadastrado neste campo na origem representou uma perda de 3.669.003 registros. Por este motivo foi escolhido considerar 1 aos valores nulos na origem ao invés de tirá-los da análise.



* + **mother\_residence\_fu:**

Coluna gerada a partir da coluna mother\_residence\_state da tabela natality. Aos valores vazios da tabela de origem serão representados como nulo nesta view.

* + **mother\_race:**

Coluna gerada a partir da coluna mother\_race da tabela natality. Quando o valor da origem for nulo, nesta view o valor será o número 9.

* + **mother\_age:**

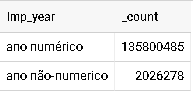
Coluna gerada a partir da coluna mother\_ age da tabela natality.

* + **gestation\_weeks:**

Coluna gerada a partir da coluna gestation\_weeks da tabela natality. Esta coluna representa a quantidade de semanas de gestação, essa é uma informação muito importante para o modelo. Quando esta coluna está nula na origem o valor dela é calculado através da diferença entre a data de nascimento (coluna birth\_fu) e a data da última menstruação (coluna lmp\_date).

* + **lmp\_year:**

Coluna gerada a partir da coluna lmp da tabela natality. Na tabela de origem o ano da última menstruação é uma parte da coluna lmp. Ele está presente nos últimos 4 digitos da coluna lmp da tabela de natality. Ao tentar extrair o ano da coluna lmp foi identificado valores não numéricos como o caractere “-”. Existem 2.026.278 registros que se encontram nessa situação.



Ao tentar identificar uma possível solução para este problema foi identificada a seguinte característica ao executar a seguinte query:

SELECT

SUBSTR(natality.lmp, 5, 4) AS lmp\_year,

FORMAT(

'%s%s-',

SUBSTR(SAFE\_CAST(natality.year AS STRING), 0, 2),

SUBSTR(SAFE\_CAST(natality.year AS STRING), 4, 1)

) AS lmp\_year\_calculated,

natality.year,

count(1) AS \_count

FROM `bigquery-public-data.samples.natality` AS natality

WHERE SAFE\_CAST(SUBSTR(natality.lmp, 5, 4) AS INT64) IS NULL

GROUP BY lmp\_year, lmp\_year\_calculated, year

E obter o seguinte resultado:

|  |  |  |  |
| --- | --- | --- | --- |
| **lmp\_year** | **lmp\_year\_calculated** | **year** | **\_count** |
| -198 | -198 | 19**7**8 | 108040 |
| -195 | -195 | 19**7**5 | 74771 |
| -199 | -199 | 19**7**9 | 88131 |
| -194 | -194 | 19**8**4 | 127188 |
| -193 | -193 | 19**7**3 | 71241 |
| -197 | -197 | 19**7**7 | 84523 |
| -192 | -192 | 19**8**2 | 129738 |
| -190 | -190 | 19**7**0 | 65772 |
| -199 | -199 | 19**6**9 | 63602 |
| -191 | -191 | 19**7**1 | 58951 |
| -198 | -198 | 19**8**8 | 144234 |
| -194 | -194 | 19**7**4 | 73714 |
| -190 | -190 | 19**8**0 | 154446 |
| -197 | -197 | 19**8**7 | 135489 |
| -196 | -196 | 19**7**6 | 76596 |
| -195 | -195 | 19**8**5 | 125009 |
| -192 | -192 | 19**7**2 | 65355 |
| -193 | -193 | 19**8**3 | 126806 |
| -191 | -191 | 19**8**1 | 120425 |
| -196 | -196 | 19**8**6 | 132247 |

Foi identificado que o caractere “-” que está presente na parte do ano da coluna lmp dos 2.026.278 registros corresponde à dezena da coluna year. A prova real foi a observação da coluna lmp\_year\_calculated que foi gerado a partir da coluna year e que bate 100% com a parte de ano da coluna lmp para os registros que possuem o caractere “-”. Desse modo, foi realizado o tratamento para obtenção do ano da coluna lmp e jogado para a coluna lmp\_year após este tratamento. Será atribuído ao lmp\_year um valor nulo caso o lmp\_year\_calculated não retorne um valor válido.

* + **lmp\_month:**

Coluna gerada a partir da coluna lmp da tabela natality. Na tabela de origem o mês da última menstruação é uma parte da coluna lmp. Ele está presente no 3º e no 4º caractere da coluna lmp da tabela de natality. Caso o valor seja 99 na tabela de origem o valor será atribuído como nulo nesta view.

* + **lmp\_day:**

Coluna gerada a partir da coluna lmp da tabela natality. Na tabela de origem o dia da última menstruação é uma parte da coluna lmp. Ele está presente nos 2 primeiros caracteres da coluna lmp da tabela de natality. Caso o valor seja 99 na tabela de origem o valor será atribuído como nulo nesta view.

* + **is\_mother\_married:**

Coluna gerada a partir da coluna mother\_married da tabela natality.

* + **mother\_birth\_fu:**

Coluna gerada a partir da coluna mother\_birth\_state da tabela natality. Aos valores vazios da tabela de origem serão representados como nulo nesta view.

* + **cigarettes\_per\_day:**

Coluna gerada a partir das colunas cigarettes\_per\_day e cigarette\_use da tabela natality. Caso a coluna cigarettes\_per\_day esteja maior que zero é utilizado o valor desta coluna, caso contrário se o valor da coluna cigarette\_use é verdadeiro, então é atribuído o valor 1, mesmo se o valor da coluna cigarettes\_per\_day for 0. Este é um campo muito importante para o modelo, desse modo foi escolhido este tratamento para tentar remover o máximo de valores faltantes possível.

* + **drinks\_per\_week:**

Coluna gerada a partir das colunas drinks\_per\_week e alcohol\_use da tabela natality. Caso a coluna drinks\_per\_week esteja maior que zero é utilizado o valor desta coluna, caso contrário se o valor da coluna alcohol\_use é verdadeiro, então é atribuído o valor 1, mesmo se o valor da coluna drinks\_per\_week for 0. Este é um campo muito importante para o modelo, desse modo foi escolhido este tratamento para tentar remover o máximo de valores faltantes possível.

* + **weight\_gain\_pounds:**

Coluna gerada a partir da coluna weight\_gain\_pounds da tabela natality.

* + **born\_alive\_alive:**

Coluna gerada a partir da coluna born\_alive\_alive da tabela natality. Na view este valor será zero caso o valor seja nulo na origem.

* + **born\_alive\_dead:**

Coluna gerada a partir da coluna born\_alive\_dead da tabela natality. Na view este valor será zero caso o valor seja nulo na origem.

* + **ever\_born:**

Coluna gerada a partir da coluna ever\_born da tabela natality. Na view este valor será zero caso o valor seja nulo na origem.

* + **father\_race:**

Coluna gerada a partir da coluna father\_race da tabela natality. Na view este valor será 9 caso o valor seja nulo na origem.

* + **father\_age:**

Coluna gerada a partir da coluna father\_age da tabela natality.

* + **is\_born\_dead:**

Coluna gerada a partir da coluna is\_born\_dead da tabela natality. Na view este valor será falso caso o valor seja nulo ou <= 0 na origem.

* **prepared\_gdl\_metrics**

Tabela que possui o IDH de cada estado concatenado a todas as demais variáveis do GDL que compõem este índice.

* + **state:**

Coluna gerada a partir da coluna state\_name da tabela fips\_codes\_states.

* + **fu:**

Coluna gerada a partir da coluna state\_postal\_abbreviation da tabela fips\_codes\_states.

* + **year:**

Coluna gerada a partir de um pivot das colunas que possuem correlação com os anos nas tabelas de origem. Ex.: colunas que vão de “\_1990” até “\_2017”. Cada coluna dessas nas tabelas de origem da Global Data Lab (GDL) viram linhas para cada estado nesta view.

* + **educational\_index:**

Coluna gerada a partir do pivot feito para as colunas de ano da tabela gdl\_educational\_index\_data.

* + **expected\_years\_schooling:**

Coluna gerada a partir do pivot feito para as colunas de ano da tabela gdl\_expected\_years\_schooling\_data.

* + **health\_index:**

Coluna gerada a partir do pivot feito para as colunas de ano da tabela gdl\_health\_index\_data.

* + **income\_index:**

Coluna gerada a partir do pivot feito para as colunas de ano da tabela gdl\_income\_index\_data.

* + **life\_expectancy:**

Coluna gerada a partir do pivot feito para as colunas de ano da tabela gdl\_life\_expectancy\_data.

* + **gross\_national\_income\_per\_capita:**

Coluna gerada a partir do pivot feito para as colunas de ano da tabela gdl\_gross\_national\_income\_per\_capita\_data.

* + **mean\_years\_schooling:**

Coluna gerada a partir do pivot feito para as colunas de ano da tabela gdl\_mean\_years\_schooling\_data.

* + **population\_size\_in\_thousands:**

Coluna gerada a partir do pivot feito para as colunas de ano da tabela gdl\_population\_size\_in\_thousands\_data.

* + **sub\_national\_hdi:**

Coluna gerada a partir do pivot feito para as colunas de ano da tabela gdl\_sub\_national\_hdi\_data.

## 3.3. model-data

As seguintes tabelas/views foram geradas a partir do cruzamento dos dados preparados e servem para treino e teste (calibragem) do modelo:

* **logistic\_regression\_is\_born\_dead\_data\_source\_2**

Esta tabela faz um cruzamento entre todos os dados trabalhados da área de stage e seleciona as variáveis que serão utilizada para calibragem do modelo.

* + **is\_male:**

Coluna gerada a partir da coluna is\_male da tabela de prepared\_natality.

* + **weight\_pounds:**

Coluna gerada a partir da coluna weight\_pounds da tabela de prepared\_natality.

* + **plurality:**

Coluna gerada a partir da coluna plurality da tabela de prepared\_natality.

* + **mother\_race:**

Coluna gerada a partir da coluna mother\_race da tabela de prepared\_natality.

* + **mother\_age:**

Coluna gerada a partir da coluna mother\_age da tabela de prepared\_natality.

* + **gestation\_weeks:**

Coluna gerada a partir da coluna gestation\_weeks da tabela de prepared\_natality.

* + **is\_mother\_married:**

Coluna gerada a partir da coluna is\_mother\_married da tabela de prepared\_natality.

* + **cigarettes\_per\_day:**

Coluna gerada a partir da coluna cigarettes\_per\_day da tabela de prepared\_natality.

* + **drinks\_per\_week:**

Coluna gerada a partir da coluna drinks\_per\_week da tabela de prepared\_natality.

* + **ever\_born:**

Coluna gerada a partir da coluna ever\_born da tabela de prepared\_natality.

* + **father\_race:**

Coluna gerada a partir da coluna father\_race da tabela de prepared\_natality.

* + **father\_age:**

Coluna gerada a partir da coluna father\_age da tabela de prepared\_natality.

* + **birth\_fu\_birth\_year\_sub\_national\_hdi:**

Coluna gerada a partir da coluna sub\_national\_hdi da tabela de prepared\_gdl\_metrics para o estado de nascimento (birth\_fu) no ano de nascimento (birth\_year) da tabela prepared\_natality.

* + **mother\_residence\_fu\_lmp\_year\_sub\_national\_hdi:**

Coluna gerada a partir da coluna sub\_national\_hdi da tabela de prepared\_gdl\_metrics para o estado de em que a mãe reside (mother\_residence\_fu) no ano da última menstruação (lmp\_year), ou seja, no início da gravidez registrado na tabela prepared\_natality.

* + **mother\_residence\_fu\_birth\_year\_sub\_national\_hdi:**

Coluna gerada a partir da coluna sub\_national\_hdi da tabela de prepared\_gdl\_metrics para o estado de em que a mãe reside (mother\_residence\_fu) no ano de nascimento (birth\_year), ou seja, no final da gravidez registrado na tabela prepared\_natality.

* + **is\_born\_dead:**

Coluna gerada a partir da coluna is\_born\_dead da tabela de prepared\_natality. Esta é a variável resposta do modelo, todas as demais são variáveis explicativas.

# 4. Análise e Exploração dos Dados

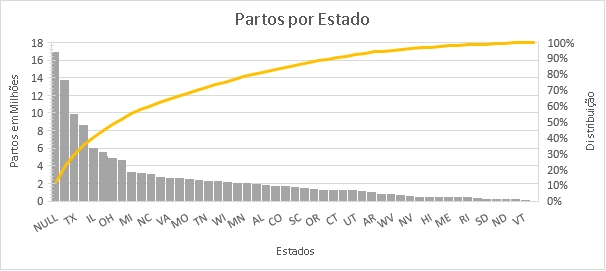
Após o tratamento dos dados nulos, vazios e inválidos das tabelas de origem (raw) na área de stage, foi feito um cruzamento obrigatório (INNER JOIN) entre a view prepared\_natality com a view prepared\_gdi\_metrics por ano e por estado (UF). Nesse cruzamento apenas ficarão disponíveis para a anáise os registros de natalidade estejam nos estados e nos anos das métricas de IDH que estão consideradas na análise. O cruzamento entre da view prepared\_natality com a view prepared\_gdi\_metrics é feito três vezes para representar os seguintes momentos:

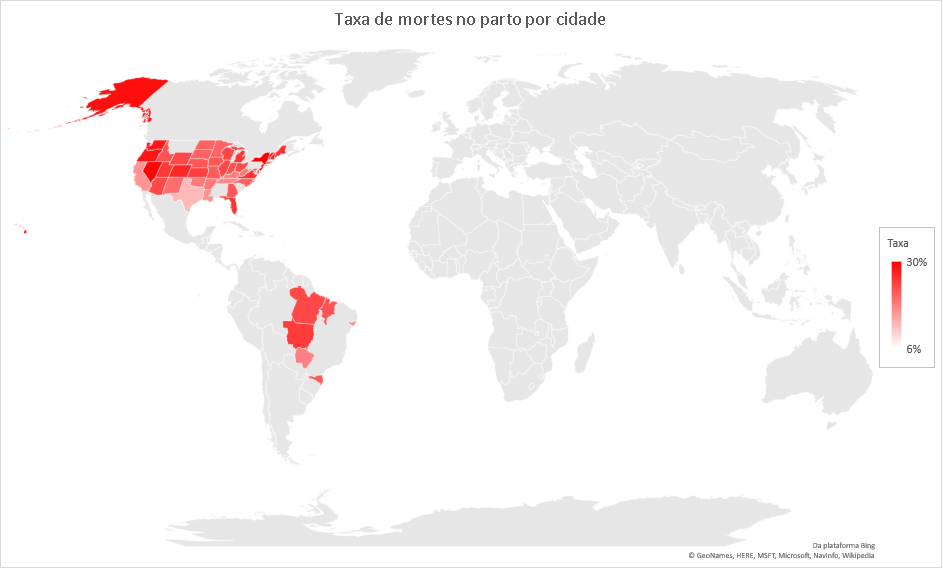
* IDH do estado e no ano de parto
* IDH do estado onde a mãe vivia no início da gestação
* IDH do estado onde a mãe vive atualmente

*Observação: Existem casos em que a mãe vive em um estado e tem filho em outro estado.*

No cruzamento também foi desconsiderado da análise registros de parto com semanas de gestação inválida (menor que 3 e maior que 42 semanas, estes valores representam o tempo mínimo e máximo de uma gravidez) a quantidade de registros caiu para 45.110.893, ou seja, houve uma redução de 67.2% de registros nesse cruzamento.

Abaixo segue alguns gráficos que representam algumas das análises exploratórias feitas:





Para maiores detalhes dos dados utilizados como base desta análise exploratória, segue link da planilha base dos relatórios acima: <https://github.com/herberton/tcc.cdbd.puc.mg/raw/master/doc/analise%20exploratoria.xlsx>

# 5. Criação de Modelos de Machine Learning

Foi feita uma análise de correlação de Pearson para todas as variáveis, esta análise de correlação resultou no seguinte relatório:

|  |  |  |
| --- | --- | --- |
| **Feature** | **Correlation** | **Adjusted Correlation Percent** |
| ever\_born | 0,312828828 | 31,28% |
| born\_alive\_alive | 0,308798897 | 30,88% |
| mother\_age | 0,189561163 | 18,96% |
| born\_alive\_dead | 0,075522229 | 7,55% |
| birth\_fu\_birth\_year\_educational\_index | 0,038041258 | 3,80% |
| mother\_residence\_fu\_birth\_yea\_educational\_index | 0,037882974 | 3,79% |
| birth\_fu\_birth\_year\_expected\_years\_schooling | 0,037728787 | 3,77% |
| birth\_fu\_birth\_year\_mean\_years\_schooling | 0,037602938 | 3,76% |
| mother\_residence\_fu\_birth\_yea\_expected\_years\_schooling | 0,037563221 | 3,76% |
| mother\_residence\_fu\_birth\_yea\_mean\_years\_schooling | 0,037455693 | 3,75% |
| father\_race\_black | 0,036736623 | 3,67% |
| mother\_residence\_fu\_lmp\_year\_educational\_index | 0,035908185 | 3,59% |
| mother\_residence\_fu\_lmp\_year\_mean\_years\_schooling | 0,035836945 | 3,58% |
| mother\_residence\_fu\_lmp\_year\_expected\_years\_schooling | 0,035380032 | 3,54% |
| is\_mother\_married\_false | -0,035185896 | 3,52% |
| is\_mother\_married\_true | 0,035185896 | 3,52% |
| birth\_fu\_birth\_year\_population\_size\_in\_thousands | -0,034597538 | 3,46% |
| mother\_residence\_fu\_birth\_yea\_population\_size\_in\_thousands | -0,034564438 | 3,46% |
| mother\_residence\_fu\_lmp\_year\_population\_size\_in\_thousands | -0,034366566 | 3,44% |
| mother\_race\_black | 0,033778855 | 3,38% |
| gestation\_weeks | -0,029116339 | 2,91% |
| Plurality | 0,024905161 | 2,49% |
| birth\_fu\_birth\_year\_sub\_national\_hdi | 0,023801494 | 2,38% |
| mother\_residence\_fu\_birth\_yea\_sub\_national\_hdi | 0,023709572 | 2,37% |
| father\_age | 0,023193478 | 2,32% |
| mother\_residence\_fu\_lmp\_year\_sub\_national\_hdi | 0,023186995 | 2,32% |
| mother\_race\_white | -0,021539372 | 2,15% |
| drinks\_per\_week | 0,016628796 | 1,66% |
| mother\_residence\_fu\_lmp\_year\_gross\_national\_income\_per\_capita | 0,014576358 | 1,46% |
| mother\_residence\_fu\_lmp\_year\_income\_index | 0,014196607 | 1,42% |
| mother\_residence\_fu\_birth\_yea\_gross\_national\_income\_per\_capita | 0,014061212 | 1,41% |
| birth\_fu\_birth\_year\_gross\_national\_income\_per\_capita | 0,013867945 | 1,39% |
| birth\_fu\_birth\_year\_income\_index | 0,013690259 | 1,37% |
| mother\_residence\_fu\_birth\_yea\_income\_index | 0,013659522 | 1,37% |
| father\_race\_unknown\_other | -0,013129664 | 1,31% |
| mother\_race\_vietnamese | -0,011757831 | 1,18% |
| mother\_race\_chinese | -0,011475241 | 1,15% |
| father\_race\_vietnamese | -0,010621731 | 1,06% |
| cigarettes\_per\_day | 0,010372811 | 1,04% |
| father\_race\_chinese | -0,010109932 | 1,01% |
| father\_race\_white | -0,009731645 | 0,97% |
| mother\_race\_american\_indian | 0,009590985 | 0,96% |
| weight\_pounds | 0,009405941 | 0,94% |
| mother\_race\_filipino | -0,00893404 | 0,89% |
| mother\_race\_asian\_indian | -0,008766364 | 0,88% |
| father\_race\_asian\_indian | -0,008405727 | 0,84% |
| father\_race\_american\_indian | 0,008032482 | 0,80% |
| father\_race\_filipino | -0,007253314 | 0,73% |
| mother\_race\_hawaiian | 0,006012812 | 0,60% |
| father\_race\_hawaiian | 0,005388207 | 0,54% |
| father\_race\_korean | -0,005152752 | 0,52% |
| mother\_race\_korean | -0,005062475 | 0,51% |
| mother\_residence\_fu\_birth\_yea\_health\_index | 0,004769535 | 0,48% |
| mother\_residence\_fu\_birth\_yea\_life\_expectancy | 0,004755459 | 0,48% |
| mother\_race\_unknown\_other | -0,004730407 | 0,47% |
| birth\_fu\_birth\_year\_health\_index | 0,004527559 | 0,45% |
| birth\_fu\_birth\_year\_life\_expectancy | 0,00451249 | 0,45% |
| mother\_residence\_fu\_lmp\_year\_life\_expectancy | 0,004458251 | 0,45% |
| mother\_residence\_fu\_lmp\_year\_health\_index | 0,004448523 | 0,44% |
| mother\_race\_japanese | -0,003146682 | 0,31% |
| is\_male\_false | 0,00179285 | 0,18% |
| is\_male\_true | -0,00179285 | 0,18% |
| father\_race\_japanese | -0,001790605 | 0,18% |

Testando as melhores combinações através das variáveis que mais tem correlação foram escolhidas as seguintes variáveis explicativas que obteve o modelo com o melhor resultado:

* is\_male
* weight\_pounds
* plurality
* mother\_race
* mother\_age
* gestation\_weeks
* is\_mother\_married
* cigarettes\_per\_day
* drinks\_per\_week
* ever\_born
* father\_race
* father\_age
* birth\_fu\_birth\_year\_sub\_national\_hdi
* mother\_residence\_fu\_lmp\_year\_sub\_national\_hdi
* mother\_residence\_fu\_birth\_year\_sub\_national\_hdi

E a seguinte variável resposta que indica se haverá morte no parto e a probabilidade este evento acontecer:

* Is\_born\_dead

Foi criado um modelo de regressão logística através da ferramenta BigQuery da Google Cloud Platform. Através do BigQuery a criação do modelo foi viabilizada através do comando SQL e englobou separação de base de treino e teste e seleção das variáveis (explicativas e resposta) referenciando a view logistic\_regression\_is\_born\_dead\_data\_source\_2 criada a partir das views da área de stage-data. O seguinte comando foi utilizado para criação do modelo de regressão logística:

CREATE OR REPLACE MODEL

`tcc-puc-mg-2019.model.logistic\_regression\_is\_born\_dead\_2`

OPTIONS (

MODEL\_TYPE = 'LOGISTIC\_REG',

AUTO\_CLASS\_WEIGHTS = TRUE,

INPUT\_LABEL\_COLS = ['is\_born\_dead']

) AS

SELECT

data\_source.is\_male,

data\_source.weight\_pounds,

data\_source.plurality,

data\_source.mother\_race,

data\_source.mother\_age,

data\_source.gestation\_weeks,

data\_source.is\_mother\_married,

data\_source.cigarettes\_per\_day,

data\_source.drinks\_per\_week,

data\_source.ever\_born,

data\_source.father\_race,

data\_source.father\_age,

data\_source.birth\_fu\_birth\_year\_sub\_national\_hdi,

data\_source.mother\_residence\_fu\_lmp\_year\_sub\_national\_hdi,

data\_source.mother\_residence\_fu\_birth\_year\_sub\_national\_hdi,

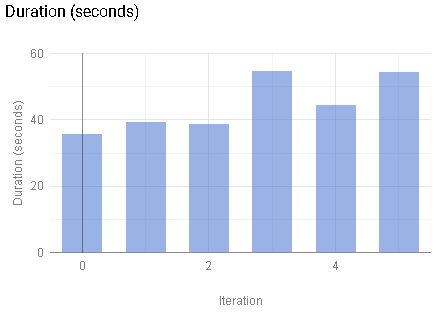
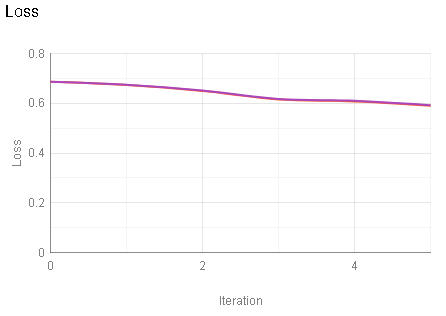
data\_source.is\_born\_dead

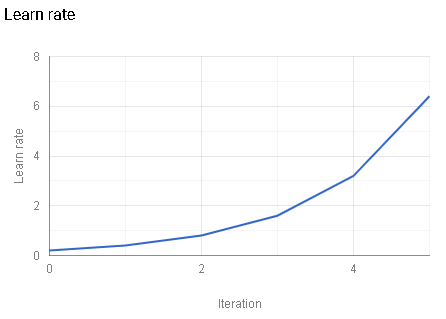
FROM

`tcc-puc-mg-2019.model.logistic\_regression\_is\_born\_dead\_data\_source\_2` AS data\_source

## 5.1. Resultados do Treino do Modelo:

Abaixo, segue gráficos referentes à forma de como o treino foi realizado (quantidade e duração das de iterações de treino, curva de aprendizagem à cada iteração etc.).





Para efetuar a predição é necessário rodar o seguinte comando:

SELECT

predicted\_is\_born\_dead,

predicted\_is\_born\_dead\_probs AS metrics

FROM

ML.PREDICT(

MODEL `tcc-puc-mg-2019.model.logistic\_regression\_is\_born\_dead\_2`,

(

SELECT

'true' AS is\_male,

6.63370946358 AS weight\_pounds,

1 AS plurality,

'Filipino' AS mother\_race,

45 AS mother\_age,

38 AS gestation\_weeks,

'true' AS is\_mother\_married,

0 AS cigarettes\_per\_day,

0 AS drinks\_per\_week,

0 AS ever\_born,

'Filipino' AS father\_race,

49 AS father\_age,

0.908 AS birth\_fu\_birth\_year\_sub\_national\_hdi,

0.905 AS mother\_residence\_fu\_lmp\_year\_sub\_national\_hdi,

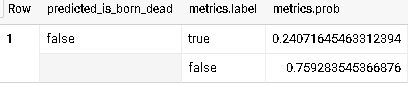
0.908 AS mother\_residence\_fu\_birth\_year\_sub\_national\_hdi

),

STRUCT(0.4422 AS threshold)

)

O resultado da execução do modelo é o seguinte:



Este resultado indica que não ocorrerá uma morte no parto com 75.92% de confiança. Este percentual de confiança foi calculado dado o “threshold” de 0.4422 que foi configurado na query de predição. Esse “threshold” é configurável nos modelos de regressão logística e pode ser adaptado à cada calibragem do modelo para alcançar a melhor predição.

Na chamada são passados 3 dados de IDH através das colunas que terminam com o sufixo “\_hdi”. Para obter os dados de IDH é necessário realizar a busca por ano e estado na view prepared\_gdl\_metrics da seguinte forma:

WITH

birth\_fu\_birth\_year AS (

SELECT sub\_national\_hdi

FROM `tcc-puc-mg-2019.stage.prepared\_gdl\_metrics`

WHERE fu = 'MS' -- estado de nascimento

AND year = 2010 /\*ano de nascimento \*/),

mother\_residence\_fu\_lmp\_year AS (

SELECT sub\_national\_hdi

FROM `tcc-puc-mg-2019.stage.prepared\_gdl\_metrics`

WHERE fu = 'MS' -- estado de residência da mãe no ano da última menstruação

AND year = 2009 /\*ano da última menstruação\*/),

mother\_residence\_fu\_birth\_year AS (

SELECT sub\_national\_hdi

FROM `tcc-puc-mg-2019.stage.prepared\_gdl\_metrics`

WHERE fu = 'MS' -- estado de residência da mãe no ano de nascimento

AND year = 2010 /\*ano de nascimento\*/ )

SELECT 'birth\_fu\_birth\_year' AS metric,

birth\_fu\_birth\_year.sub\_national\_hdi

FROM birth\_fu\_birth\_year

UNION ALL

SELECT 'mother\_residence\_fu\_lmp\_year' AS metric,

mother\_residence\_fu\_lmp\_year.sub\_national\_hdi

FROM mother\_residence\_fu\_lmp\_year

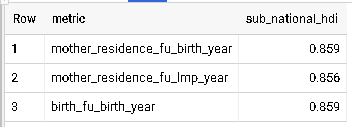
UNION ALL

SELECT 'mother\_residence\_fu\_birth\_year' AS metric,

mother\_residence\_fu\_birth\_year.sub\_national\_hdi

FROM mother\_residence\_fu\_birth\_year

A query acima retorna o seguinte resultado que pode ser utilizado para o modelo preditivo:



Para os dados de raça da mãe e do pai devem ser um destes listados abaixo:

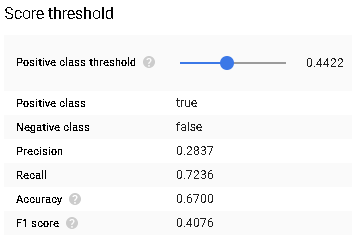
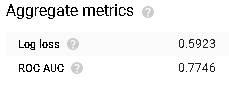


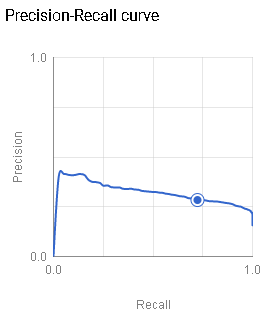
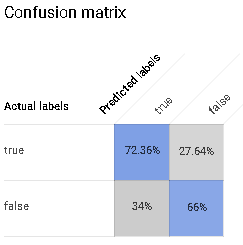
# 6. Apresentação dos Resultados

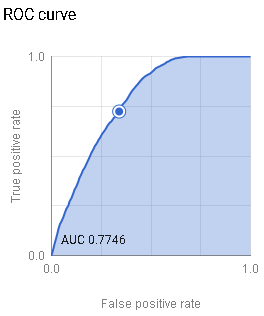
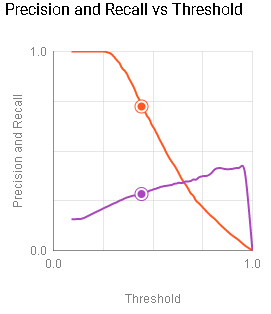
Abaixo seguem dados estatísticos de avaliação do modelo gerado.

## 6.1. Métricas de Qualidade do Modelo

Abaixo seguem algumas métricas sobre a qualidade do modelo de regressão logística gerado:







# 7. Links

* Projeto no Github: <https://github.com/herberton/tcc.cdbd.puc.mg>
* Apresentação do Projeto no YouTube: <https://youtu.be/RRn1ohAGJ10>

# REFERÊNCIAS

NAVLANI, Avinash. **Understanding Logistic Regression in Python.** Internet: Site da Data Camp, 2018. Link: <https://www.datacamp.com/community/tutorials/understanding-logistic-regression-python>. Acessado em 06 de outubro de 2019.

UNDP, United Nations Development Programme. **Human Development – Indices and Indicators.** New York: Site da UNDP, 2018. Link: <http://www.hdr.undp.org/sites/default/files/2018_human_development_statistical_update.pdf>. Acessado em 08 de outubro de 2019.

UNDP, United Nations Development Programme. **Technical notes.** New York: Site da UNDP, 2018. Link: <http://hdr.undp.org/sites/default/files/hdr2018_technical_notes.pdf>

SMITS, Jeroen & PERMANYER, Iñaki. **Construction of the Sub-National Human Development Index.** Internet: Site da Global Data Lab, 2019. Link: <https://globaldatalab.org/shdi/about/>. Acessado em 25 de setembro de 2019.

NTS, Núcleo de Telessaúde Sergipe. **Qual o período limite de uma gestação? Houve alguma alteração recente?** Internet: Site da Biblioteca Virtual em Saúde da Atenção Primária à Saúde (BVS APS), 2014. Link: <https://aps.bvs.br/aps/qual-o-periodo-limite-de-uma-gestacao-houve-alguma-alteracao-recente/>. Acessado em 19 de setembro de 2019.