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

1.1. 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.

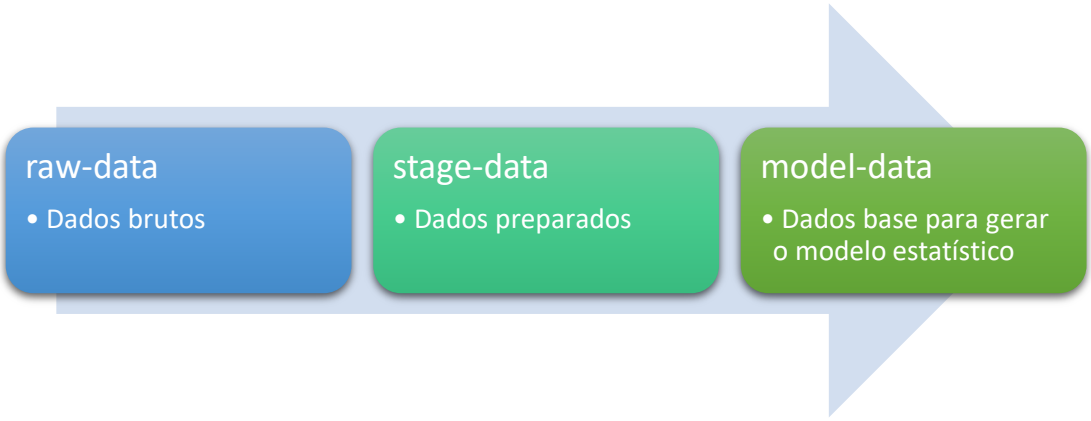
2. Coleta de Dados

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.

plurality	_count
não-nulos	134157760
nulos	3669003

- **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.

lmp_year	_count
ano numérico	135800485
ano não-numerico	2026278

  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	1978	108040
-195	-195	1975	74771
-199	-199	1979	88131
-194	-194	1984	127188
-193	-193	1973	71241
-197	-197	1977	84523
-192	-192	1982	129738
-190	-190	1970	65772
-199	-199	1969	63602
-191	-191	1971	58951
-198	-198	1988	144234
-194	-194	1974	73714
-190	-190	1980	154446
-197	-197	1987	135489
-196	-196	1976	76596
-195	-195	1985	125009
-192	-192	1972	65355
-193	-193	1983	126806
-191	-191	1981	120425
-196	-196	1986	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

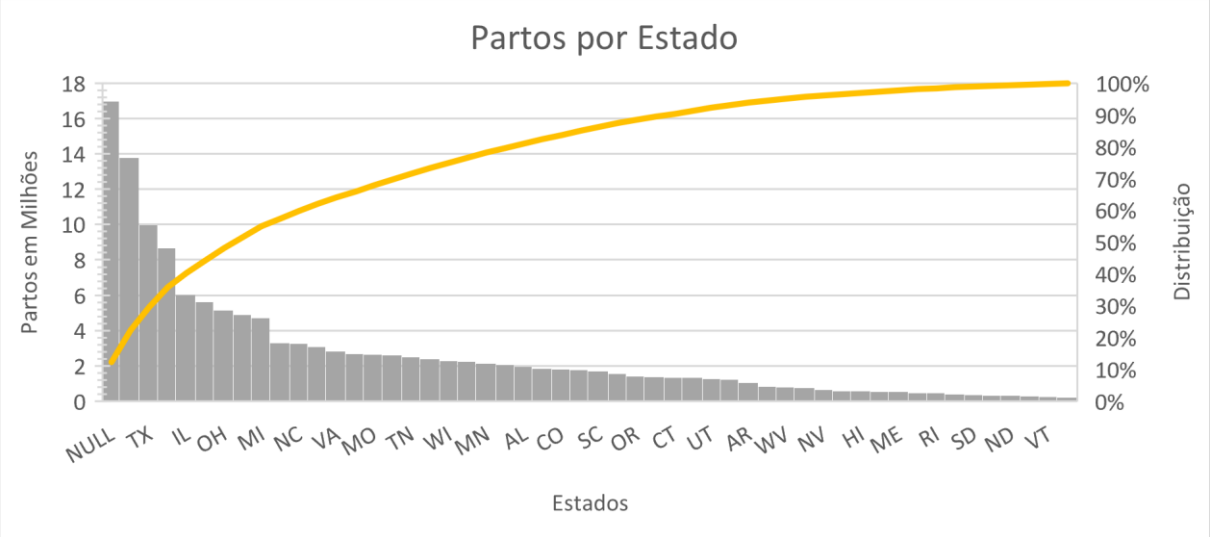
Após o tratamento dos dados nulos, vazios e inválidos das tabelas de origem (raw) na área de stage, foi feito um cruzamento obrigatório (INNER JOIN) entre a view prepared_natality com a view prepared_gdi_metrics por ano e por estado (UF). Nesse cruzamento apenas ficarão disponíveis para a anáise os registros 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:

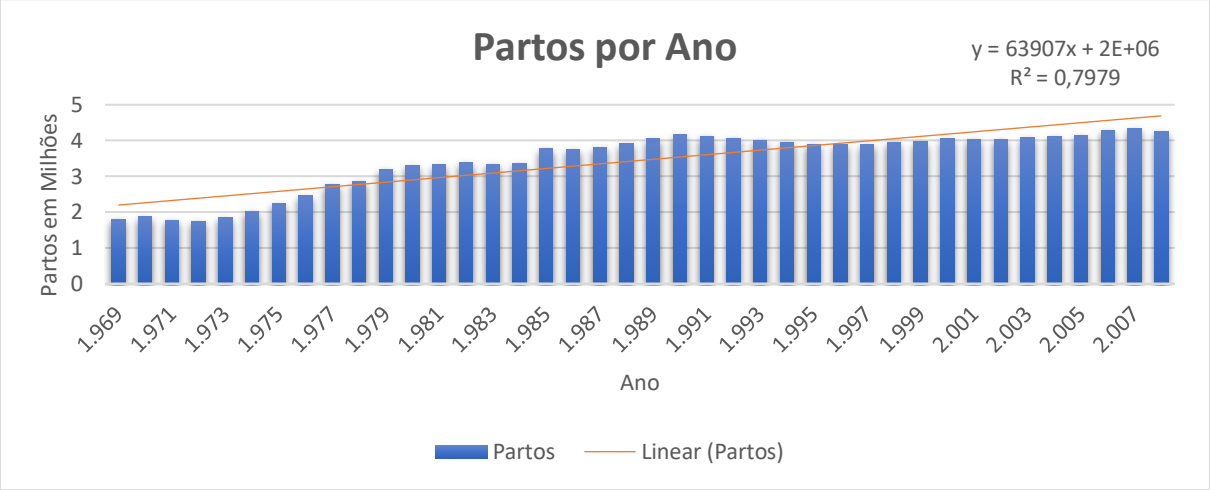
- IDH do estado e no ano de parto
- IDH do estado onde a mãe vivia no início da gestação
- IDH do estado onde a mãe vive atualmente

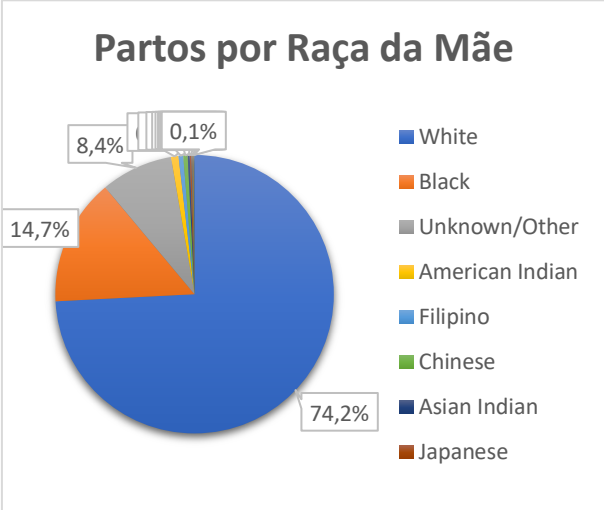
  Observação: Existem casos em que a mãe vive em um estado e tem filho em outro estado.

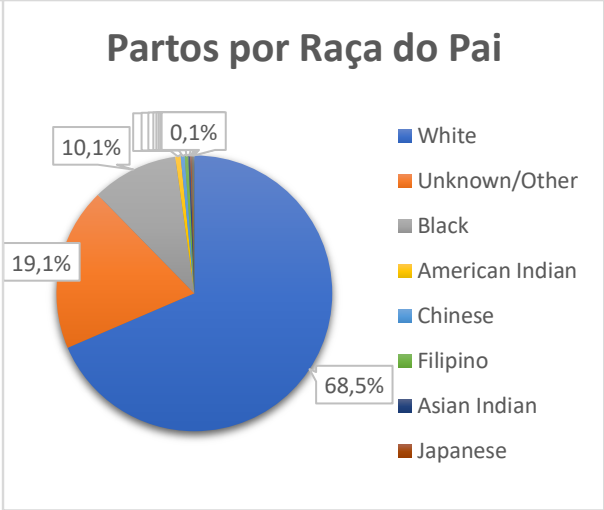
No cruzamento também foi desconsiderado da análise registros de parto com semanas de gestação inválida (menor que 3 e maior que 42 semanas, estes valores representam o tempo mínimo e máximo de uma gravidez) a quantidade de registros caiu para 45.110.893, ou seja, houve uma redução de 67.2% de registros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

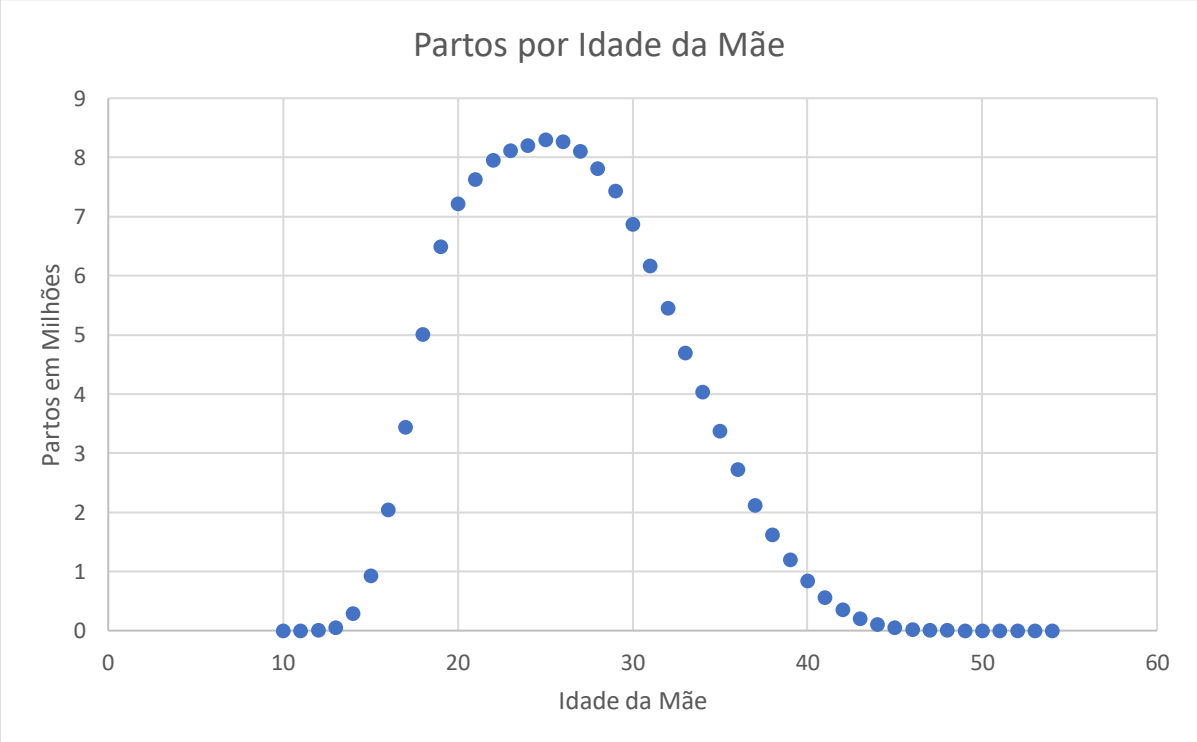
Abaixo segue alguns gráficos que representam algumas das análises exploratórias feitas:



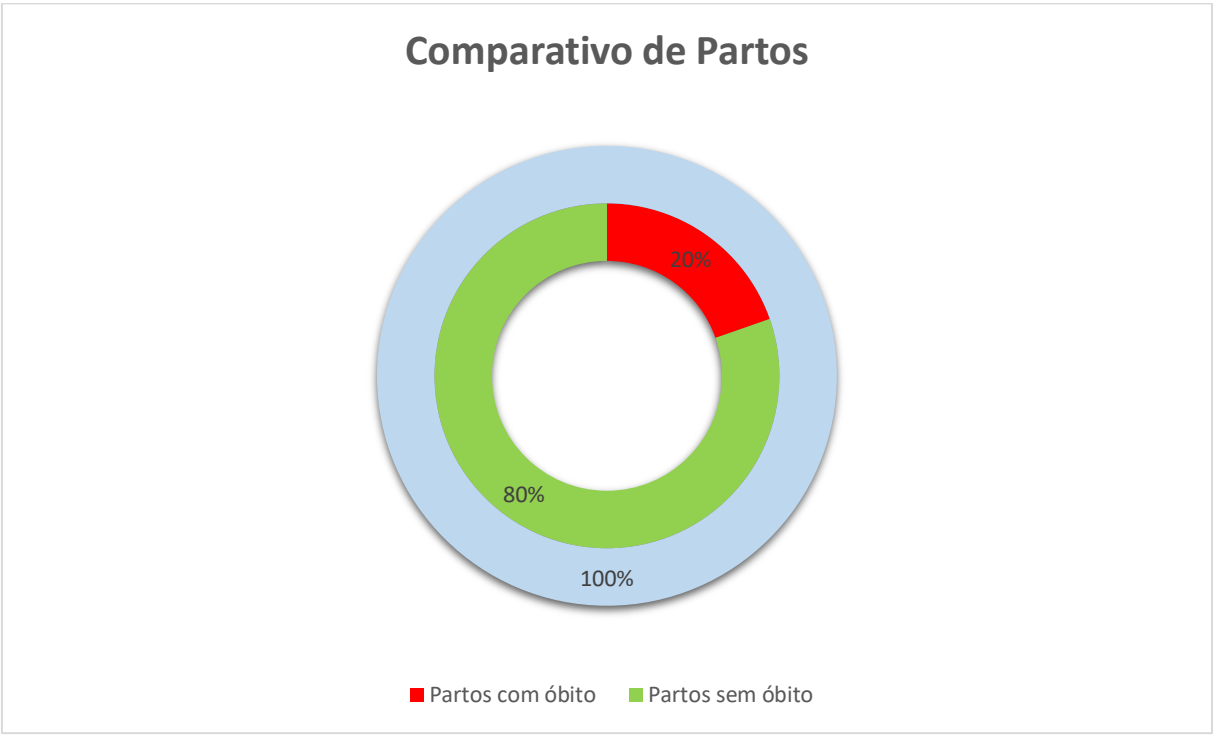

Partos por Ano
y = 63907x + 2E+06
R² = 0,7979
Partos em Milhões
0
1
2
3
4
5
1.969
1.971
1.973
1.975
1.977
1.979
1.981
1.983
1.985
1.987
1.989
1.991
1.993
1.995
1.997
1.999
2.001
2.003
2.005
2.007
Ano
Partos
Linear (Partos)


Partos por Raça da Mãe
8,4%
0,1%
14,7%
74,2%
White
Black
Unknown/Other
American Indian
Filipino
Chinese
Asian Indian
Japanese

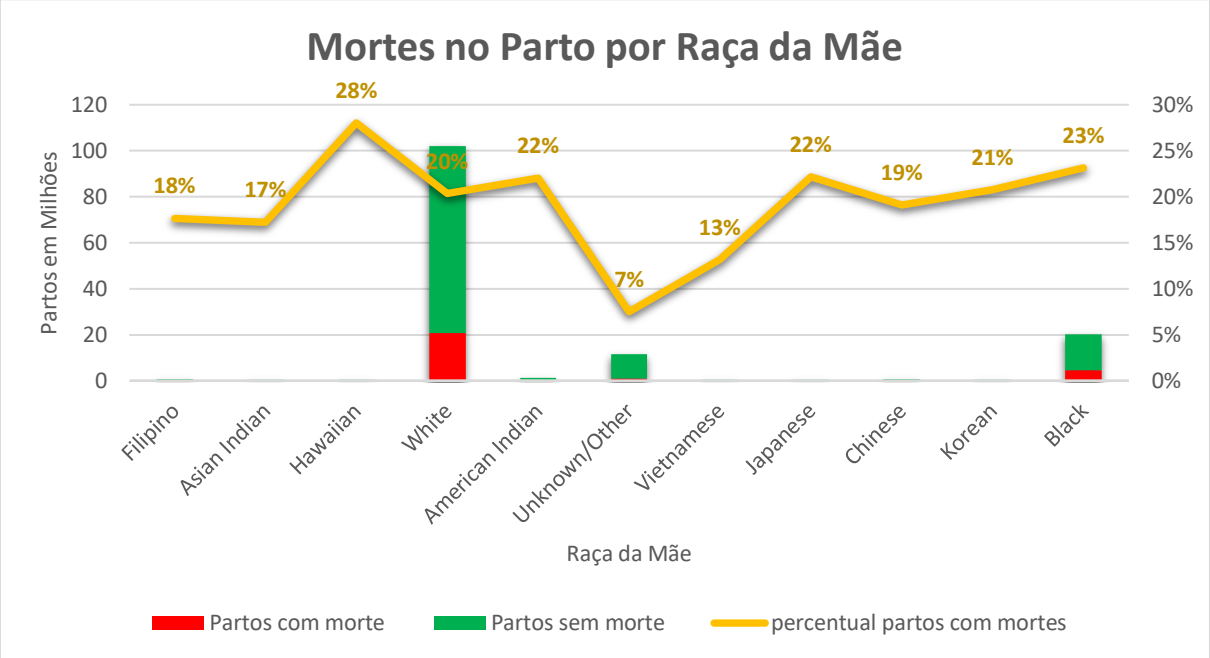

Partos por Raça do Pai
10,1%
0,1%
19,1%
68,5%
White
Unknown/Other
Black
American Indian
Chinese
Filipino
Asian Indian
Japanese

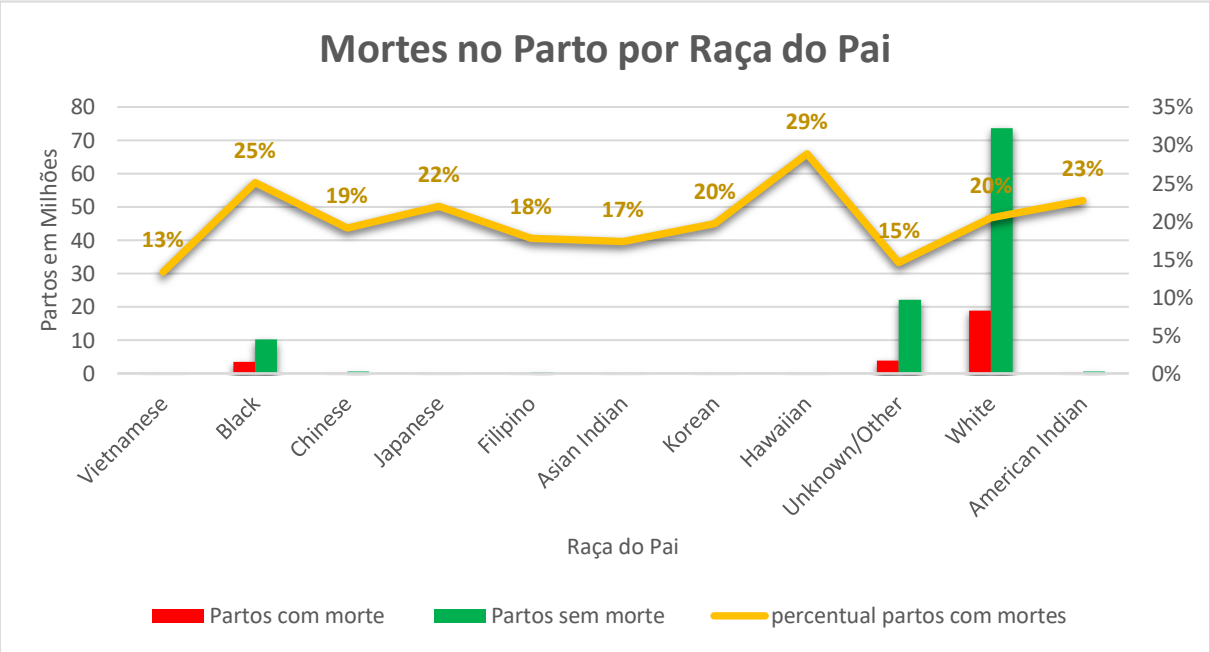

Partos por Idade da Mãe
Partos em Milhões
0
1
2
3
4
5
6
7
8
9
0
10
20
30
40
50
60
Idade da Mãe


Partos por Idade do Pai
Partos em Milhões
0
1
2
3
4
5
6
7
8
0
20
40
60
80
100
120
Idade do Pai


Comparativo de Partos
20%
80%
100%
Partos com óbito
Partos sem óbito


Taxa de mortes no parto por cidade
Taxa
30%
6%
Da plataforma Bing
© GeoNames, HERE, MSFT, Microsoft, Navlnfo, Wikipedia


Mortes no Parto por Raça da Mãe
Partos em Milhões
120
100
80
60
40
20
0
30%
25%
20%
15%
10%
5%
0%
18%
17%
28%
20%
22%
7%
13%
22%
19%
21%
23%
Filipino
Asian Indian
Hawaiian
White
American Indian
Unknown/Other
Vietnamese
Japanese
Chinese
Korean
Black
Raça da Mãe
Partos com morte
Partos sem morte
percentual partos com mortes


Mortes no Parto por Raça do Pai
Partos em Milhões
80
70
60
50
40
30
20
10
0
35%
30%
25%
20%
15%
10%
5%
0%
13%
25%
19%
22%
18%
17%
20%
29%
15%
20%
23%
Vietnamese
Black
Chinese
Japanese
Filipino
Asian Indian
Korean
Hawaiian
Unknown/Other
White
American Indian
Raça do Pai
Partos com morte
Partos sem morte
percentual partos com mortes

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:

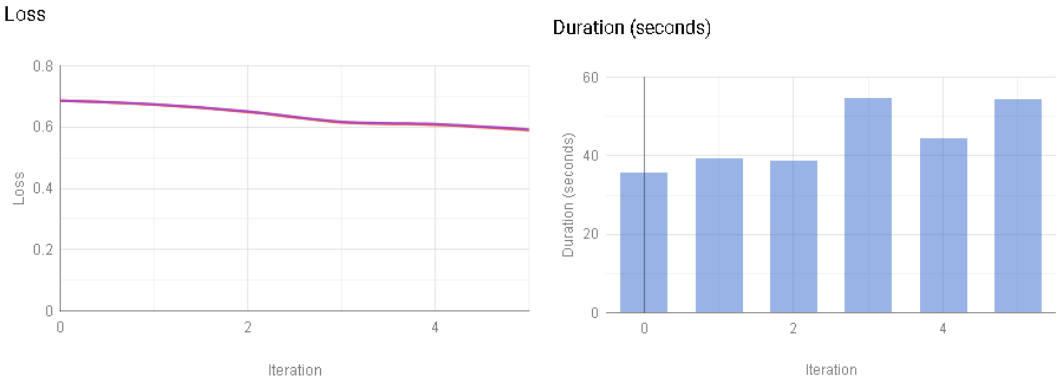
- Is_born_dead

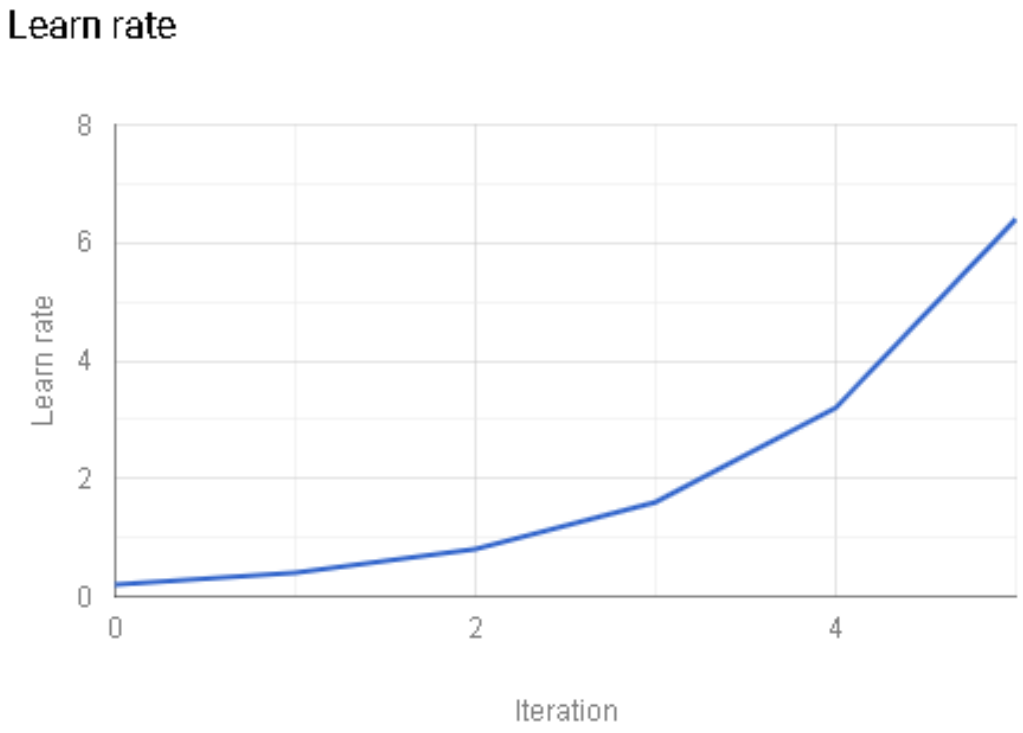
Foi criado um modelo de regressão logística através da ferramenta BigQuery da Google Cloud Platform. Através do BigQuery a criação do modelo foi viabilizada através do comando SQL e englobou separação de base de treino e teste e seleção das variáveis (explicativas e resposta) referenciando a view logistic_regression_is_born_dead_data_source_2 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:

Row	predicted_is_born_dead	metrics.label	metrics.prob
1	false	true	0.24071645463312394
		false	0.759283545366876

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:

Row	metric	sub_national_hdi
1	mother_residence_fu_birth_year	0.859
2	mother_residence_fu_lmp_year	0.856
3	birth_fu_birth_year	0.859

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

Row	race
1	American Indian
2	Asian Indian
3	Black
4	Chinese
5	Filipino
6	Hawaiian
7	Japanese
8	Korean
9	Unknown/Other
10	Vietnamese
11	White

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:

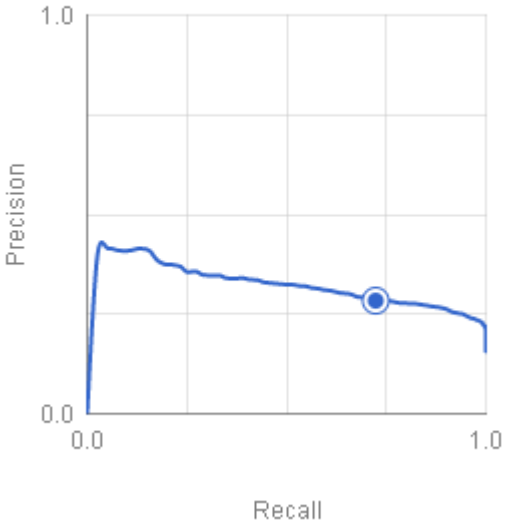
Score threshold

Positive class threshold	0.4422
Positive class	true
Negative class	false
Precision	0.2837
Recall	0.7236
Accuracy	0.6700
F1 score	0.4076

Aggregate metrics

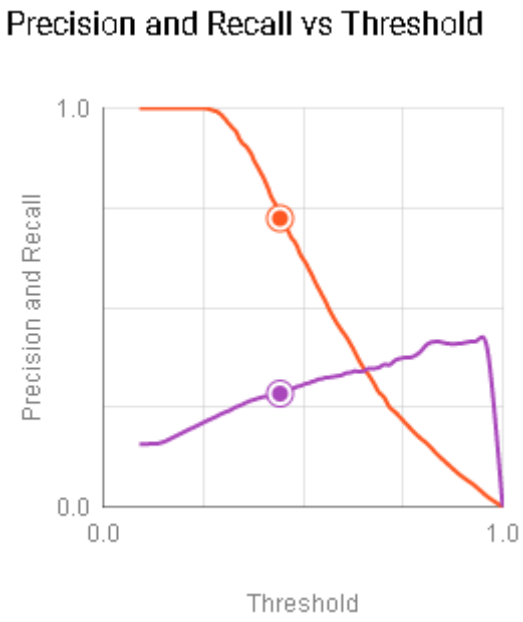
Log loss	0.5923
ROC AUC	0.7746

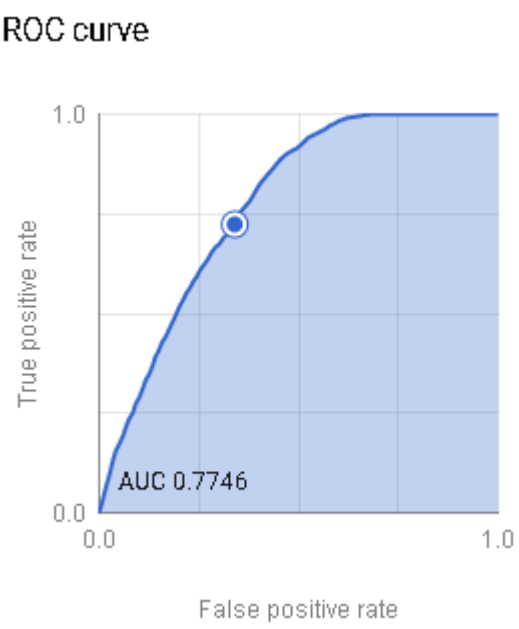
Precision-Recall curve



Confusion matrix

Actual labels \ Predicted labels	true	false
true	72.36%	27.64%
false	34%	66%


Precision and Recall vs Threshold
1.0
0.0
Precision and Recall
0.0
1.0
Threshold


ROC curve
1.0
0.0
True positive rate
AUC 0.7746
0.0
1.0
False positive rate

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.