Machine Learning Engineer Nanodegree

Projeto final - Projeto Capstone

Helton Souza Lima

Junho, 2019

I. Definição

Visão geral do projeto

O Programa Bolsa Família (PBF) é o maior programa de distribuição de renda do Brasil [1], através de um benefício em dinheiro transferido diretamente do governo federal para famílias dentro da linha da pobreza e extrema pobreza, para garantir um alívio mais imediato à pobreza, complementando a renda dessas famílias e condicionando à participação nos serviços de saúde e educação. De acordo com o artigo da Dra. Daniela Dias Kuhn [2], o programa foi efetivo na melhoria dos índices de desenvolvimento humano no Estado do Rio Grande do Sul. Podemos citar outro estudo, realizado em Minas Gerais [3] que aponta a mesma conclusão no âmbito deste estado.

Por outro lado, é recorrente a veiculação de notícias [4] referentes a fraudes nos benefícios do Programa Bolsa Família. Essas fraudes acarretam saques de valores superiores ao necessário para o atingimento do objetivo do programa e precisam ser eliminadas, pois acarretam um custo desnecessário ao governo, chegando ao patamar de bilhões [5] de reais.

A empresa em que trabalho é a DATAPREV [6], empresa de processamento de dados do governo federal. Uma atividade recorrente de nossa empresa é o levantamento e cruzamento de informações entre bases de dados para verificar o correto cumprimento de políticas públicas através de sistemas informatizados. O trabalho com os dados do Bolsa Família permitirá a investigação de situações semelhantes a outras que fazem parte das recorrentes demandas dentro da empresa, e a experiência poderá ser útil dentro de um contexto semelhante ao problema abordado neste trabalho.

Descrição do problema

O público-alvo do PBF são as pessoas que estão dentro da faixa da pobreza ou pobreza extrema. Entende-se que os volumes financeiros disponibilizados para o programa é proporcional à quantidade de pessoas dentro das faixas sociais que são alvo do programa, de forma que, a partir de dados de informações sociais e econômicas, como a população total, esperança de vida ao nascer, taxa de analfabetismo, percentual de crianças na escola, taxa de frequência, renda per capita, percentual de distribuição de renda, proporção de pobres, etc, é possível predizer o volume financeiro a ser utilizado para o PBF. Em suma, este trabalho visa verificar se municípios com índices mais baixos (índices que compõem o IDH) recebem mais recursos do PBF, pois correspondem a municípios mais pobres. De forma análoga, em tese, municípios com índices mais altos recebem menos recursos do PBF, considerando a quantidade de pessoas residentes nesses municípios.

Uma das respostas que se desejou responder foi: Será que existem municípios com alto IDHM mas que, mesmo assim, recebem muitos recursos do PBF, em comparação com outros municípios semelhantes?
Sendo assim, de posse dos dados granularizados a nível de município brasileiro, relativos à pesquisa de mapeamento do Índice de Desenvolvimento Humano Municipal (IDHM) no ano de 2010, utilizou-se modelos de machine learning que foram treinados utilizando-se os dados de parte desses municípios e foram capazes de predizer o volume financeiro da outra parte desses municípios. Em um momento inicial, a análise dos dados apontou a correlação entre os indicadores sociais e o volume financeiro do PBF associado com cada município. Em seguida, foi possível identificar alguns municípios que apontaram discrepância nessa correlação e foram apontados como municípios onde é possível que tenha sofrido uma maior influência de fraudes.

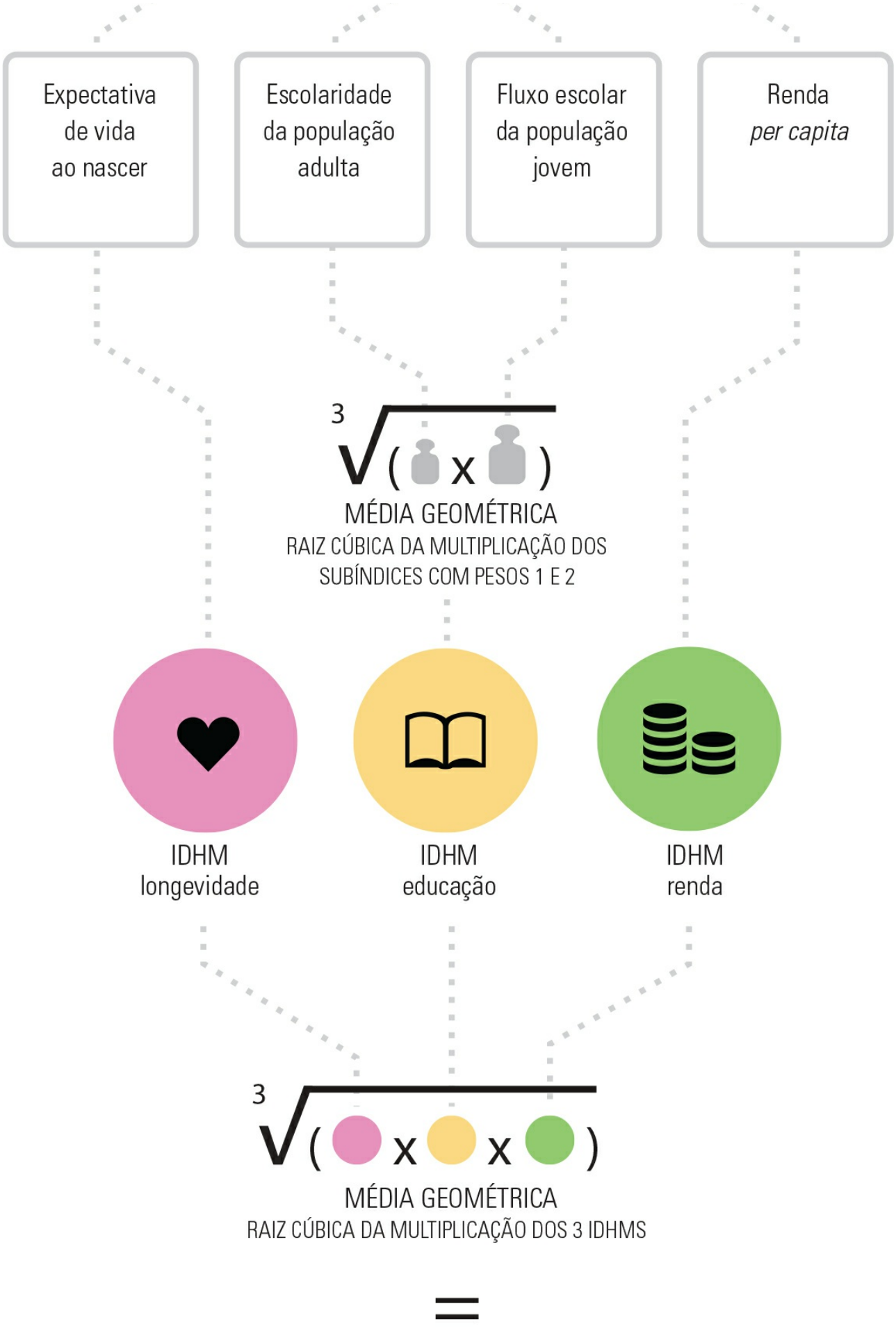
Métricas

O valor a ser previsto é um valor contínuo, correspondente ao valor, em reais, que é disponibilizado para ser sacado pelos beneficiários do Bolsa Família para cada município. A métrica que foi utilizada é o Root Mean Squared Erros (RMSE), pois é uma métrica que avalia a distância entre o valor previsto e o valor real. Essa métrica é calculada pelo próprio scikit-learn comparando os valores previstos e os valores reais, através da utilização do método "score" dos modelos de regressão. O RMSE é definido como (1 - u/v), onde u é soma das diferenças ao quadrado (quadrado(real - previsto)).sum() e v é o total da soma dos quadrados (quadrado(real - média(real))).sum(). A melhor possibilidade é o valor de "score" ser 1.0 e pode ser negativo se o modelo se comportou de forma muito ruim [11].

II. Análise

Exploração dos dados

Os dados utilizados foram obtidos de duas fontes. A primeira fonte são os dados relacionados ao Índice de Desenvolvimento Humano Municipal (IDHM), disponibilizado pelo site Atlas do Desenvolvimento Humano no Brasil [7] ou no site da Kaggle [8]. Os dados do IDHM são disponibilizados para cada um dos 5565 municípios brasileiros, sendo composto por dados que podem ser agrupados em 3 dimensões: dados sobre longevidade, dados sobre o nível de acesso ao conhecimento e dados sobre a renda. O cálculo do IDHM foi realizado a partir das informações dos 3 últimos Censos Demográficos do IBGE (1991, 2000 e 2010). Neste trabalho foram utilizados os dados do IDHM de 2010.


Expectativa de vida ao nascer
Escolaridade da população adulta
Fluxo escolar da população jovem
Renda *per capita*
3
MÉDIA GEOMÉTRICA
RAIZ CÚBICA DA MULTIPLICAÇÃO DOS SUBÍNDICES COM PESOS 1 E 2
IDHM longevidade
IDHM educação
IDHM renda
3
MÉDIA GEOMÉTRICA
RAIZ CÚBICA DA MULTIPLICAÇÃO DOS 3 IDHMS
=

IDHM

A segunda fonte são os dados relacionados à quantidade de famílias beneficiárias e o total de pagamentos disponibilizados pelo PBF para cada município brasileiro. Os dados são disponibilizados pelo Ministério da Cidadania [9]. Os dados utilizados são de janeiro de 2010, ou seja, 7 anos após o ano de lançamento do PBF, que pode ser considerado como suficiente para o programa ter atingido uma maturidade em sua operacionalização e gestão e os dados serem considerados consolidados. Também são dados que coincidem com o ano da realização do Censo, em 2010, como forma de aproximar o levantamento social realizado pelo Censo dos dados de recursos disponibilizados pelo Bolsa Família.

União dos dados

A primeira etapa do trabalho foi a união de ambas as fontes de dados para formar um único conjunto de dados. O resultado final é composto de 5565 linhas (correspondentes a cada município) e 241 colunas (4 colunas dos dados do Bolsa Família, incluindo a *Quantidade de Famílias Beneficiárias do Bolsa Família* e o *Valor Repassado para Bolsa Família*, e 237 colunas dos dados para composição do IDHM).

Tratamento de variáveis categóricas

- Foram identificadas 7 colunas com valores não-numéricos e de códigos pertencentes a domínios:
    - **ano**: Sempre o mesmo ano em todas as linhas (2010)
    - **codmun6, ibge, codmun7**: Códigos identificadores do município
    - **município**: Nome do município
    - **anomes**: Competência (mês + ano) do valor diponibilizado pelo Bolsa Família, sempre com valor "201001", que significa janeiro de 2010.
    - **uf**: Código do IBGE identificador da Unidade de Federação ao qual o município pertence.
- Todas essas variáveis foram removidas para a continuação da análise exploratória e alimentação dos modelos de predição.

Transformação de valores

Através da exploração inicial dos dados, verificou-se que a variável **idhm** possuía alguns registros entre 0 e 1 e o restante dos registros entre 400 e 900. Em verificações individuais destes casos, percebeu-se que os registros estavam apenas transformados para valores entre 0 e 1. Por exemplo, para o município de Cabixi, em Rondônia, o valor que se verificou foi 0,65. Entretanto, após pesquisa no portal Atlas Brasil, este município foi avaliado com IDHM 650.

Portanto, decidiu-se realizar a transformação destes casos para que todos ficassem com a mesma base. O mesmo procedimento foi realizado para **idhm_e, idhm_l, idhm_r, i_freq_prop e i_escolaridade**.

Dados ausentes e outliers

Não foram identificados dados ausentes no conjunto de dados, após a realização de busca por lacunas. Em relação aos *outliers*, nenhum caso foi interpretado como *outlier*. Todas as variáveis analisadas individualmente apresentaram distribuição normal ou distribuição normal mista, com dois picos. Foram avaliadas individualmente as variáveis **valor_repassado_bolsa_familia, qtd_familias_beneficiarias_bolsa_familia, idhm, idhm_e, idhm_l, idhm_r, i_freq_prop, i_escolaridade, theil, gini, pmpob, pind e pesotot**.

Visualização Exploratória

Variáveis avaliadas individualmente através de gráficos

Conforme relatado na seção anterior, as variáveis avaliadas individualmente foram analisadas através de gráficos que estão a seguir. A escolha das variáveis analisadas individualmente nesta fase de exploração dos dados foi apenas baseado no sentimento de importância das variáveis, dadas as informações obtidas no atlasbrasil.org.br :

valor_repassado_bolsa_familia

- Grande parte dos municípios recebe até 100 mil reais. A quantidade de municípios com valor maior que 200 mil reais está em torno de 25%
-

qtd_familias_beneficiarias_bolsa_familia

- Grande parte dos municípios possui até mil famílias beneficiadas. A quantidade de municípios com mais de 2 mil famílias beneficiárias está em torno de 25%. O maior valor é 181531 família beneficiárias.
-

idhm

- Índice de Desenvolvimento Humano do Município. É uma distribuição normal mista, com dois pontos de picos, próximo dos valores 600 e 720. O menor valor é 418 e o maior é 862.
-

gini

- Mede o grau de desigualdade existente na distribuição de indivíduos segundo a renda domiciliar per capita. Seu valor varia de 0, quando não há desigualdade (a renda domiciliar per capita de todos os indivíduos tem o mesmo valor), a 1, quando a desigualdade é máxima (apenas um indivíduo detém toda a renda).O universo de indivíduos é limitado àqueles que vivem em domicílios particulares permanentes.
-

pmpob

- Proporção dos indivíduos com renda domiciliar per capita igual ou inferior a R$ 140,00 mensais, em reais de agosto de 2010. O universo de indivíduos é limitado àqueles que vivem em domicílios particulares permanentes.

- 

pind

- Proporção dos indivíduos com renda domiciliar per capita igual ou inferior a R$ 70,00 mensais, em reais de agosto de 2010. O universo de indivíduos é limitado àqueles que vivem em domicílios particulares permanentes.
-

pesotot

- População total de cada município
- 

Análise de variáveis correlacionadas

Gráfico de correlação

- Foram utilizados gráficos de correlação para avaliar se haviam variáveis com forte correlação e pudessem ser eliminadas do modelo sem perda de informação relevante para a fase de predição.
- 

Algoritmos e técnicas

Para este trabalho, foi tomado como premissa que existe uma relação linear entre as variáveis que compõem o IDHM e o valor repassado para o Bolsa Família. Ou seja, quanto mais baixo o IDHM, maior é o valor proporcional à população do repasse de verbas referente ao Bolsa Família. Portanto, as características deste problema apontam que existem variáveis dependentes de forma linear às variáveis independentes. Sobre a existência de outliers, não foram identificados casos que pudessem se caracterizados como tal, mesmo considerando as grandes capitais brasileiras em que o volume de recursos do Bolsa Família é bem maior do que a grande maioria dos outros municípios. Entende-se que há uma relação linear com a população residente em cada município.

Os valores que se deseja predizer já existem dentro do conjunto de dados, de forma que foi possível calcular a acurácia dos algoritmos escolhidos. Todos os algoritmos escolhidos são para problemas de regressão, pois a intenção é predizer valores contínuos.

Regressão Linear

Este é o modelo mais usado para soluções lineares [12] por sua simplicidade e performance na fase de treinamento e predição. A desvantagem desse algoritmo é a sua sensibilidade em relação a outliers, caso existam.

Árvore de decisão

Modelo baseado em árvore e são fáceis de entender e visualizar [13]. Suporta variáveis categóricas ou numéricas e é capaz de resolver problemas com múltiplas saídas. Entre as desvantagens estão a criação de árvores que levem ao *overfitting* e no caso de pequenas

variações nos dados de entrada é possível que a árvore gerada mude bastante assim como as predições realizadas.

Floresta aleatória

É um modelo composto em que múltiplas árvores de decisão são combinadas para um modelo mais robusto [14], com maior acurácia e imune a sobre-ajustes. Entre as desvantagens estão menor performance quando a floresta cresce e menor entendimento sobre as suas predições.

Huber Regressor

É um modelo de regressão linear, porém mais imune a *outliers* [15].

Linear Support Vector Machine

É um modelo que se comporta bem com um número grande de variáveis porém com amostra pequena [14]. Entre as desvantagens está sua complexidade e performance que degrada muito quando a amostra aumenta.

Modelo de referência

- O modelo utilizado como referência foi o de **Regressão Linear**, por ser o mais utilizado para este tipo de problema e tem boa performance na fase de treinamento e predição [12].
- Não encontramos algum trabalho que realizou trabalho semelhante para que possamos realizar uma comparação direta.

III. Metodologia

Pré-processamento dos dados

Foram realizados diretamente no Excel a união dos dados e a transformação de valores para uma mesma base, conforme relatado em sessão anterior.

Em seguida, o que mais chamou a atenção na análise exploratória dos dados foi a quantidade de variáveis existentes no conjunto de dados referentes ao cálculo do IDHM: 237 variáveis. Uma hipótese levantada no início do trabalho e que norteou a preparação dos dados foi a possibilidade de eliminar variáveis que fossem redundantes para alimentação de modelos de machine learning. Sendo assim, como primeiro passo foram eliminadas as variáveis categóricas e, em seguida, aquelas com forte relação e que agregariam muito pouco aos modelos em relação à capacidade de predição, sendo apenas informações que deixam o processamento mais lento.

Para a identificação da relação entre as variáveis, foram utilizados gráficos de correlação. Cada gráfico conseguiu exibir a correlação de aproximadamente 90 variáveis (quando temos 237). Por isso, foi necessária a renderização de 8 gráficos com eliminações sucessivas. No total, foram eliminadas 76 variáveis, restando 161 variáveis para alimentar os modelos. Adicionalmente foi utilizado o **SelectKBest** como algoritmo de verificação das variáveis mais significativas, afim de evitar eliminações que viessem a prejudicar a predição dos modelos de aprendizado. Mesmo assim, foi observado, através do *ScatterPlot* que mesmo entre as variáveis com maior peso, havia correlação muito forte entre elas, sendo possível ainda mais eliminações.

Por fim, foram verificados que não havia nenhuma variável com valores nulos.

Implementação

Para o treinamento dos modelos, foi realizado a separação dos dados em conjunto de treinamento e conjunto de teste, através do método *train_test_split* da biblioteca Scikit-Learn. Como esta separação pode ser randômica a cada execução, foi utilizado o parâmetro **random_state** para não haver mudanças nesta separação a cada execução. Percebeu-se que, com diversas separações diferentes, alguns modelos tiveram *scores* razoavelmente diferentes entre si.

Todos os algoritmos apresentados em sessão anterior foram usados de forma idêntica, com os mesmos dados de treinamento e teste. O *score* foi impresso, assim como um gráfico exibindo os valores previstos em relação aos valores reais, como a figura a seguir:

- A proximidade da linha vermelha indica a proximidade da predição em relação ao dado real
- 

Refinamento

Para cada algoritmo utilizado para predição, foi utilizado GridSearch [17] de forma a verificar se a mudança dos hiper-parâmetros poderiam melhorar o *score* do modelo.

IV. Resultados

Avaliação e validação do modelo

Regressão Linear

O score atingido com os parâmetros-padrão foi **0,920**. Não houve melhora após a utilização do GridSearch.

Árvore de decisão

O score atingido com os parâmetros-padrão foi **0,887**. Não houve melhora após a utilização do GridSearch, que precisou de bastante tempo para completar.

Floresta aleatória

O score atingido com os parâmetros-padrão foi **0,934**. Não houve melhora após a utilização do GridSearch.

Huber Regressor

O score atingido com os parâmetros-padrão foi **0,920**. Não houve melhora após a utilização do GridSearch.

Linear Support Vector Machine

O score atingido com os parâmetros-padrão foi **0,894**. Após a utilização do GridSearch o score atingido foi **0,900**.

Justificativa

O modelo com maior *score* foi o modelo de **Floresta aleatória** (*Random Forest*), sendo um pouco melhor que o modelo de referência **Regressão Linear** (*Linear Regression*). A sua performance na fase de treinamento teve um desempenho pior, porém aceitável. Para o problema elencado neste trabalho, *Random Forest* mostra-se um dos melhores candidatos.

V. Conclusão

Visualização de forma livre

Os dois gráficos resultantes dos modelos de Regressão Linear e Floresta aleatória são exibidos abaixo em que o eixo y corresponde ao valor previsto e o eixo x corresponde ao valor real. A linha vermelha indica quando os dois valores seriam exatamente os mesmos.

- Regressão Linear
- 
- Floresta Aleatória
- 

Reflexão

Aperfeiçoamento

Referências

[1] Portal do Programa Bolsa Família. Ministério da Cidadania. (http://mds.gov.br/assuntos/bolsa-familia)

[2] Kuhn, Daniela Dias. Tonetto, Elci da Silva. O Programa Bolsa Família e os indicadores sociais no Rio Grande do Sul. Desenvolvimento em Questão (https://www.revistas.unijui.edu.br/index.php/desenvolvimentoemquestao/article/view/5799/530

[3] Denubila, Lais Atanaka. Ferreira, Marco Aurelio Marques. Monteiro, Doraliza Auxiliadora Abranches. Programa Bolsa Família: Análise Da Trajetória Dos Indicadores Sociais Em Minas Gerais. Associação Nacional de Pós-Graduação e Pesquisa em Administração (http://www.anpad.org.br/admin/pdf/apb1239.pdf)

[4] Busca no Google sobre fraudes no Bolsa Família (https://www.google.com/search?q=bolsa+fam%C3%ADlia+fraudes&rlz=1C1GCEU_pt-brBR835BR835&source=lnms&tbm=nws&sa=X&ved=0ahUKEwiz_MzgsLbhAhU7KLkGHcQzCmg

[5] "Controladoria-Geral acha R$ 1,3 bi em fraudes no Bolsa Família", Revista Exame Online, 4 de janeiro de 2018 (https://exame.abril.com.br/brasil/controladoria-geral-acha-r-13-bi-em-fraudes-no-bolsa-familia/)

[6] Portal da Dataprev. Empresa de Tecnologia e Informações da Previdência Social (http://www.dataprev.gov.br/)

[7] Portal do Atlas do Desenvolvimento Humano no Brasil (http://www.atlasbrasil.org.br/2013/pt/o_atlas/idhm/)

[8] Human Development Indexes and Census data for Brazilian municipalities. Portal Kaggle (https://www.kaggle.com/pauloeduneves/hdi-brazil-idh-brasil)

[9] Visualizador de Dados Sociais. Um portal do Ministério da Cidadania (https://aplicacoes.mds.gov.br/sagi/vis/data/data-table.php)

[10] Human Development Indexes and Census data for Brazilian municipalities. Kaggle DataSet. Setembro/2018 (https://www.kaggle.com/kerneler/starter-hdi-brazil-idh-brasil-80f68b4b-6)

[11] Método "score" do modelo Linear Regression. Biblioteca scikit-learn v0.21.2 (https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html#sklearn.linear_

[12] Comparative Study on Classic Machine learning Algorithms. Danny Varghese. Portal TowardsDataScience.com (https://towardsdatascience.com/comparative-study-on-classic-machine-learning-algorithms-24f9ff6ab222)

[13] Decision Trees. Biblioteca scikit-learn v0.21.2 (https://scikit-learn.org/stable/modules/tree.html#tree)

[14] Comparative Study on Classic Machine learning Algorithms - Part 2. Danny Varghese. Portal TowardsDataScience.com (https://medium.com/@dannymvarghese/comparative-study-on-classic-machine-learning-algorithms-part-2-5ab58b683ec0)

[15] Huber Regressor. Biblioteca scikit-learn v0.21.2 (https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.HuberRegressor.html#sklearn.linear_n

[16] Select K-Best. Biblioteca scikit-learn v0.21.2 (https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.SelectKBest.html#sklearn.feature

[17] GridSearchCV. Biblioteca scikit-learn v0.21.2 (https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html#sklearn.model_