Análise de taxas de suicídio no Brasil e no mundo

Bryan Valeriano[1], **Marcelo de Souza Pena**[1],
Marcos Freitas Parra[1], **Mauro Mascarenhas de Araújo**[1],
Rafael A. Zanatta[1], **Vinícius Lourenço da Silva**[1]

[1]Universidade Federal do ABC (UFABC)
CEP 09210-580 – Santo André – SP – Brazil

{bryan.valeriano, marcelo.pena, marcos.parra}@aluno.ufabc.edu.br

{mauro.mascarenhas, rafael.zanatta, vinicius.lourenco}@aluno.ufabc.edu.br

Abstract. *This project sought to apply data mining and analysis techniques and the regression tree with cross validation learning method to obtain the best analysis parameters for the database in order to verify the suicide rate in Brazil. After performing the analysis, it was observed that the suicide rate in Brazil increased linearly over the years, where this rate is approximately four times higher among men than women, and that GDP per capita is directly related to the increase. Suicide is a delicate and complex issue. Therefore, this study is not intended to be definitive, but rather to encourage discussion and to point out trends which may shape new studies in the area.*

Resumo. *Este projeto buscou aplicar técnicas de análise e mineração de dados utilizando o método árvore de regressão com validação cruzada para obter os melhores parâmetros para análise da base de dados a fim de verificar a taxa de suicídios no Brasil. Após executar a análise pôde-se observar que a taxa de suicídio no Brasil cresceu linearmente ao longo dos anos, sendo que ela é aproximadamente quatro vezes maior entre homens em relação às mulheres e que o PIB per capita está diretamente relacionado com o aumento das taxas de suicídio. Suicídio é um tema complexo e delicado. Portanto, este trabalho não tem como objetivo ser definitivo, mas de fomentar a discussão e apontar tendências para que novos estudos sejam realizados.*

1. Introdução

Todo suicídio é uma tragédia, sendo uma das principais causas de morte prematura no mundo. De acordo com estimativas da Organização Mundial da Saúde, mais de 800.000 pessoas morrem por suicídio a cada ano [15]. Isso corresponde a uma taxa de suicídio padronizada por idade de cerca de 11,5 por 100.000 pessoas. Esse número preocupa, pois, equivale a uma pessoa suicidando-se a cada 40 segundos, sendo que a taxa de tentativa de suicídio é três vezes maior em mulheres e a taxa de suicídio consumado é quatro vezes maior entre os homens [2]. Contudo, suicídios são evitáveis com intervenções oportunas baseadas em evidências.

O uso de dados estatísticos para identificar possíveis grupos de risco já foi usada na literatura diversas vezes [4, 8], e já foi provado, que esses dados colaboram com programas de prevenção para este tipo de fatalidade [13]. O uso de mineração de dados, também já foi empregado com resultados positivos em alguns casos, como que decisão tomar com pacientes sobreviventes de uma tentativa de suicídio, e o uso de dados de redes sociais para determinar possíveis suicídios de adolescentes Coreanos [3, 20].

Existem diferentes técnicas de mineração de dados possíveis para a análise realizada, dentre elas, podemos citar *K-vizinhos-mais-próximos*, *regressão logística*, *árvores de regressão* e o *SVM* [10]. Essas técnicas enquadram-se em técnicas de aprendizado supervisionado, onde temos dados previamente classificados, e desejamos classificar novos dados através destes [12]. Geralmente, esses métodos possuem duas etapas: uma de treinamento, na qual separamos um conjunto de dados dados de treinamento e dados de teste. Durante a etapa de treinamento, usamos apenas o conjunto determinado para treinamento, sem nunca usar os dados de teste para nenhuma parte desta etapa. Através dos dados de treinamento e do modelo escolhido, é definido um classificador, que após avaliado com alguma técnica de validação, como, por exemplo, validação cruzada, é definido qual será a classificação para novos dados [1]. Após, o classificador é avaliado com o conjunto de teste, e verificada se a porcentagem de acerto está dentro do esperado para o problema abordado [12].

Dentre os modelos de aprendizado supervisionado, optamos por *árvores de regressão*, pois são mais aconselháveis quando temos um conjunto de dados de tamanho predeterminado, portanto, neste estudo o método escolhido foi a *Decision Tree Regression* [16]. A utilização dessa técnica permite que os dados utilizados no treinamento da árvore possuam erros e também apresentem atributos com valores desconhecidos [7].

2. Objetivo

A fim de compreender melhor as estatísticas, este estudo busca identificar possíveis semelhanças em dados de pessoas com o risco de suicídio, para auxiliar na elaboração

de políticas públicas e ações preventivas.

3. Materiais e Métodos

3.1. Descrição dos dados

O banco utilizado foi gerado a partir dos dados obtidos através do site *Kaggle*, *disponível em: https://www.kaggle.com/russellyates88/suicide-rates-overview-1985-to-2016*.

O banco de dados é uma compilação de quatro outras fontes:

- United Nations Development Program. (2018). Human development index (HDI). Obtido de *http://hdr.undp.org/en/indicators/137506*
- World Bank. (2018). World development indicators: GDP (current US$) by country:1985 to 2016. Disponível em: *http://databank.worldbank.org/data/source/world-development-indicators#*
- Suicide in the Twenty-First Century [dataset]. Disponível em: *https://www.kaggle.com/szamil/suicide-in-the-twenty-first-century/notebook*
- World Health Organization. (2018). Suicide prevention. Disponível em: *http://www.who.int/mental_health/suicide-prevention/en/*

3.1.1. Tipos de dados

O dataset utilizado possui 11 atributos, sendo que, *country*, *sex*, *age*, *generation*, *continent* são categóricos, com destaque para o atributo *sex* que é uma variável categórica binária e os atributos *year*, *suicides_no*, *population*, *suicides/100k pop*, *gdp_for_year$* e *gdp_per_capita$* são variáveis numéricas.

No total, os dados possuem 27820 linhas e 11 colunas.

3.2. Métodos utilizados

3.2.1. Limpeza e preparação dos dados

Primeiramente, efetuou-se a leitura do banco de dados com auxílio da biblioteca *pandas* [14] e as linhas que continham valores nulos foram removidas. Logo após, como desejava-se analisar apenas os dados do Brasil, também com auxílio da biblioteca *pandas*, os dados foram agrupados por país e apenas os relacionados ao Brasil foram extraídos para um novo *dataframe*. Além disso, para verificar a distribuição do número de suicídios por faixa etária, os dados também foram agrupados por idade.

Em seguida, foram criadas duas funções a fim de facilitar a análise descritiva dos dados, sendo elas:

- **suicidiosPorAno:** Plota um gráfico da soma de suicídios por 100 mil habitantes em cada um dos anos disponíveis.

- **suicidiosPorIdade:** Plota gráficos da soma de suicídios por 100 mil habitantes por faixa etária em cada um dos anos.

A fim de verificar possíveis diferenças entre o número de suicídios em relação ao sexo dos indivíduos, o *dataframe* contendo os dados apenas do Brasil foram divididos por sexo em outros dois *dataframes*.

Como o algoritmo utilizado durante este trabalho não aceita variáveis categóricas foi necessário utilizar a técnica de *encoding* para transformar as variáveis categóricas em numéricas. Os atributos foram codificados utilizando as regras abaixo:

- **Sexo:** Mulher = 0, Homem = 1;

- **Faixa etária:** "5-14 years"= 0, "15-24 years"= 1, "25-34 years"= 2, "35-54 years"= 3, "55-74 years"= 4 , "75+ years"= 5;

- **Geração:** "Silent"= 0, "Generation X"= 1, "G.I Generation"= 2, "Boomers"= 3, "Generation Z"= 4, "Millenials"= 5

Por conseguinte, por tratar-se da análise de apenas um país, as colunas *"country"* e *"continent"* foram excluídas, visto que não acrescentariam nenhuma informação à analise, uma vez que todos os valores são idênticos. Além disso, a coluna *"year"* também foi excluída, pois o objetivo deste projeto é avaliar tendências atemporais de suicídios. A coluna *"population"* também foi removida, uma vez que, com o aumento da população, a taxa de suicídios também aumenta, portanto, este não seria um bom estimador para detectar os grupos de risco de suicídio.

3.2.2. Algoritmos de Aprendizado de Máquina

Os algoritmos utilizados durante o projeto foram obtidos do pacote *scikitlearn* disponíveis para Python [16].

O algoritmo utilizado para gerar os grupos de risco foi o de Árvores de decisão para regressão [5]. Os melhores parâmetros para a construção da árvore foram obtidos com a utilização do algoritmo de *Grid Search* com auxílio de validação cruzada para validação [18]. O *GridSearch* procura exaustivamente a melhor combinação entre todos os parâmetros. Os valores utilizados na busca foram:

- **"max_depth"**: 1-30;

- **"min_samples_leaf"**: 1-60;
- **"min_samples_split"**: 2-20.

A melhor combinação obtida foi a seguinte:

- **"max_depth"** = 7
- **"min_samples_leaf"** = 1
- **"min_samples_split"** = 2

Em seguida, foi utilizado o algoritmo de *Recursive Feature Elimination* [9] com auxílio de validação cruzada para avaliação de desempenho para verificar qual o número ótimo de variáveis a serem utilizadas e quais são elas.

A fim de verificar a ordem de importância das variáveis, foi utilizado o algoritmo *SelectKbest* [19] sendo que o parâmetro utilizado para escolha das variáveis foi a função estatística F-test [17]. O número K de variáveis foi escolhido com base no valor retornado pelo algoritmo de *Recursive Feature Elimination*.

4. Resultados

A seguir, pode-se observar os gráficos obtidos através dos dados.

Pode-se notar que o número de suicídios aumentou ao longo dos anos. Apesar de a taxa por 100 mil habitantes estar próxima de 1 no período em que mais houveram suicídios, houve um aumento expressivo em relação ao período inicial de coleta dos dados, em 1985.

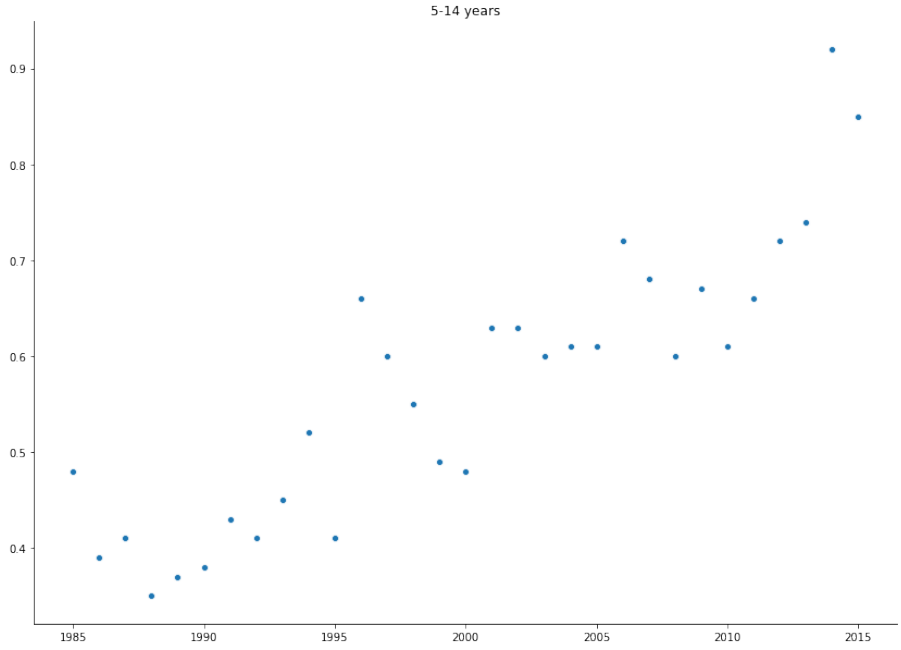

Figura 1. Suicídios no Brasil entre pessoas de 5 a 14 anos

De maneira semelhante ao gráfico de suicídios por 100 mil habitantes por ano de pessoas na faixa etária de 5 a 14 anos, também houve um aumento de ocorrências para a faixa de 15 a 24, sendo mais expressivo que na anterior, uma vez que no ano em que mais houveram suicídios, a taxa era de aproximadamente 10 ocorrências a cada 100 mil habitantes.

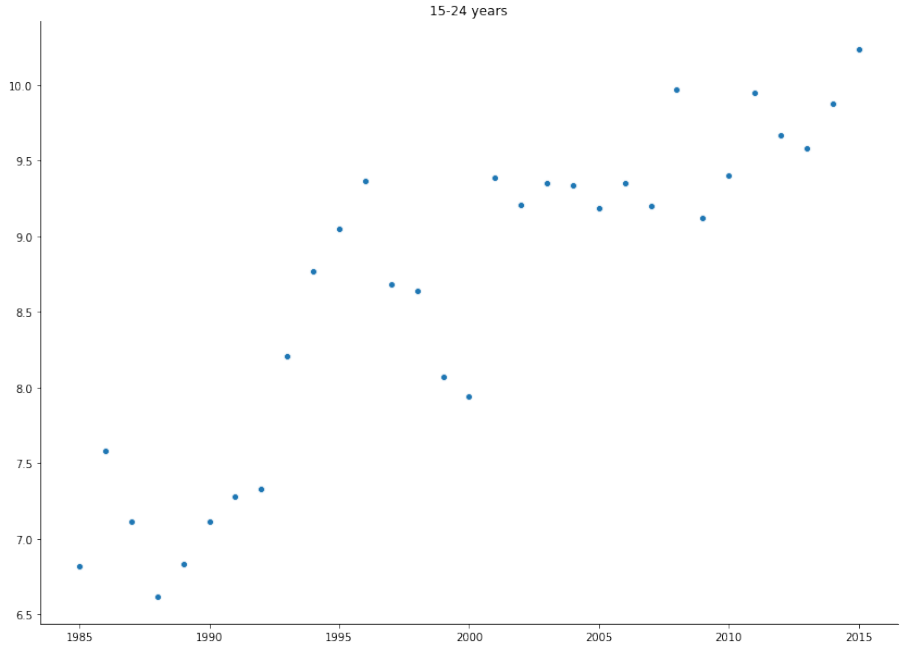

Figura 2. Suicídios no Brasil entre pessoas de 15 a 24 anos

Seguindo o padrão dos demais gráficos, também observa-se uma tendência no aumento do número de suicídios entre pessoas com idade variando de 25 a 34 anos, onde quando houve a menor ocorrência de suicídios para esta faixa etária, foi justamente próximo do ano em que mais houveram suicídios dentro da faixa etária anterior (15 a 24 anos). No ano em que houveram mais suicídios dentro dessa faixa, a taxa foi de aproximadamente 14 suicídios por 100 mil habitantes.

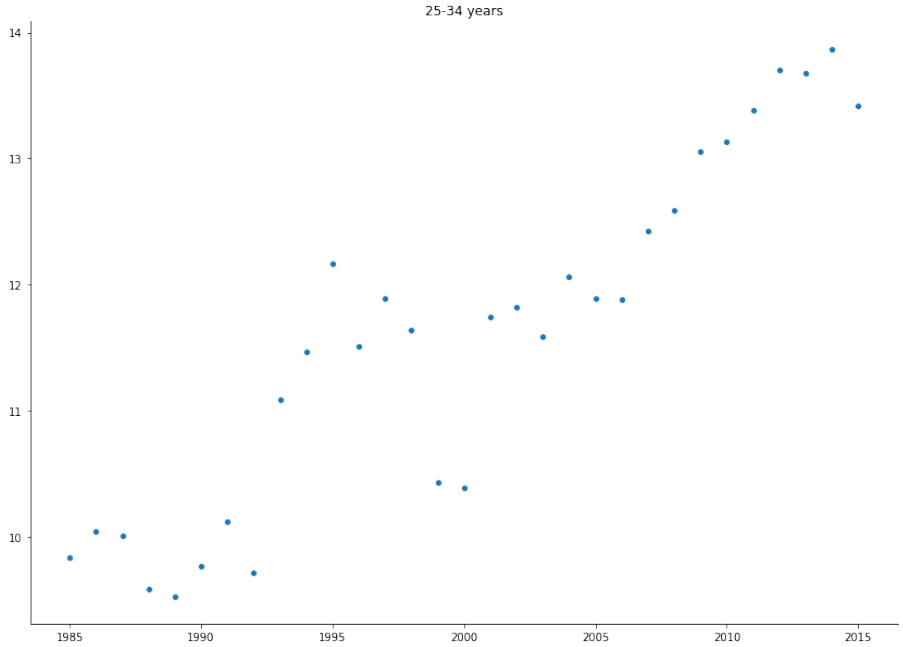

Figura 3. Suicídios no Brasil entre pessoas de 25 a 34 anos

O número de ocorrências dentro da faixa de 35-54 anos também segue uma tendência de crescimento linear, tendo números próximos aos da faixa etária anterior. Dado que, no ano em que a taxa foi maior, ela ficou próxima de 15 suicídios por 100 mil habitantes.

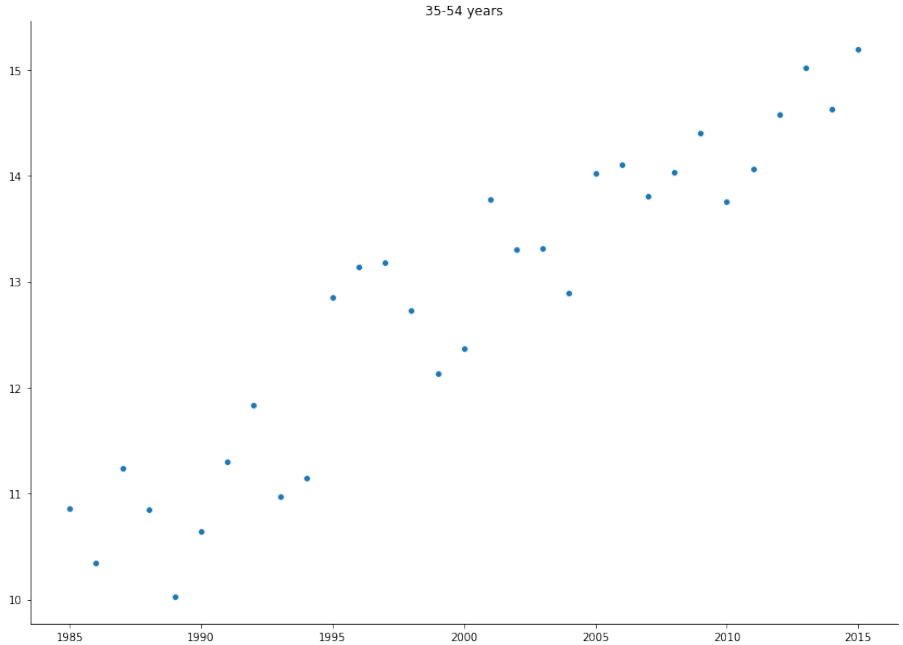

Figura 4. Suicídios no Brasil entre pessoas de 35 a 54 anos

O número de suicídios na faixa etária de 55 a 74 anos também apresenta uma

tendência de crescimento linear. Entretanto, ela possui uma variância maior em relação às demais distribuições. A maior taxa para esta faixa está por volta de 16 suicídios a cada 100 mil habitantes. A amplitude da taxa encontrada nesta faixa etária aproxima-se bastante das últimas três, porém acaba sendo um pouco maior.

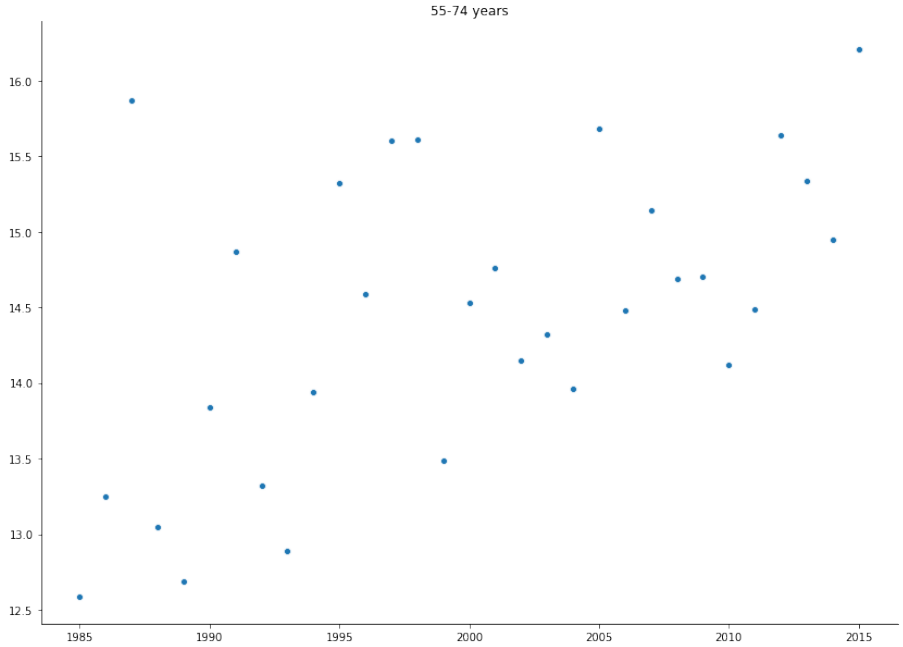

Figura 5. Suicídios no Brasil entre pessoas de 55 a 74 anos

Indo na contramão das faixas etárias anteriores, a taxa de suicídio entre pessoas com 75 ou mais está numa tendência de decrescimento linear. Entretanto, nos últimos anos, ela ficou por volta de 18-20 suicídios por 100 mil habitantes, sendo, portanto, a faixa etária que, ainda assim, mais comete suicídios em relação às demais.

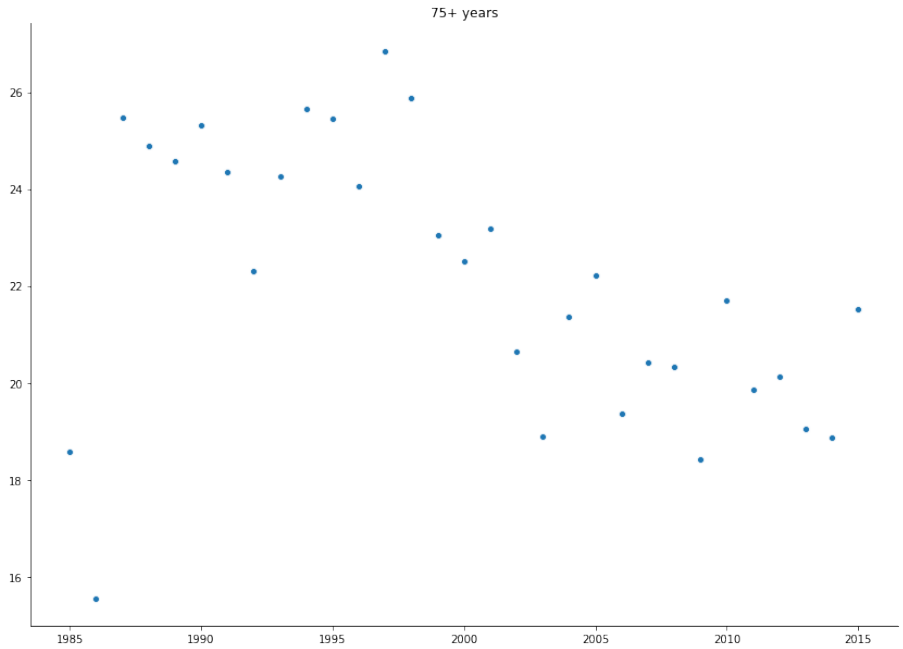

Figura 6. Suicídios no Brasil entre pessoas de 75 anos ou mais

A taxa de suicídios no Brasil como um todo (i.e, considerando todas as faixas etárias) também segue um crescimento linear, sendo que a menor taxa foi de 57 suicídios por 100 mil habitantes e a maior taxa foi de 77 suicídios por 100 mil habitantes. Entretanto, até meados dos anos 1995-1999 havia uma crescente mais acentuada, que teve sua inclinação alterada por volta dos anos 2000.

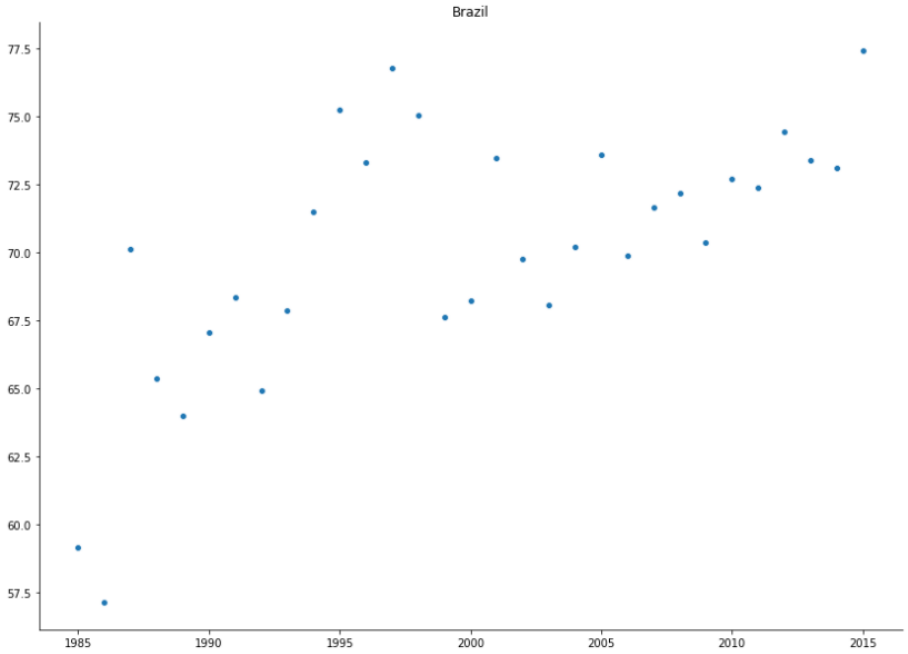

Figura 7. Suicídios no Brasil por ano

Pode-se observar no gráfico abaixo a relação entre PIB per capita e o número de suicídios por 100 mil habitantes. Aparentemente, conforme o PIB per capita aumenta, também há um aumento do número de suicídios. Porém, a mediana continua basicamente na mesma faixa de valores.

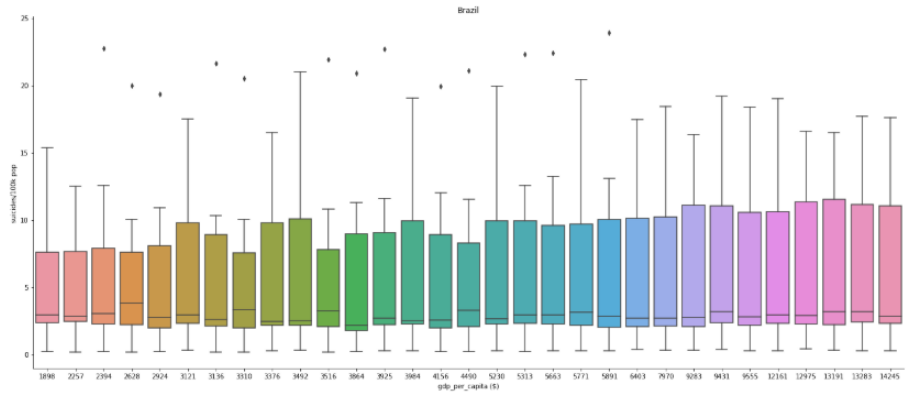

Figura 8. Relação entre suicídios e PIB per capita

No gráfico de separação entre a distribuição entre a quantidade de suicídio cometidos por homens e por mulheres, podemos observar que os homens se concentram numa faixa com amplitude maior e, com números maiores de suicídios, enquanto as mulheres tem uma frequência de suicídios concentrados na faixa dos 2000 suicídios. De fato, o número de suicídios entre os homens é maior: a soma do número de ocorrências entre os homens foi de 177.598 durante todo o período disponível nos dados, enquanto a quantidade de suicídios cometidos por mulheres foi de 49.015.

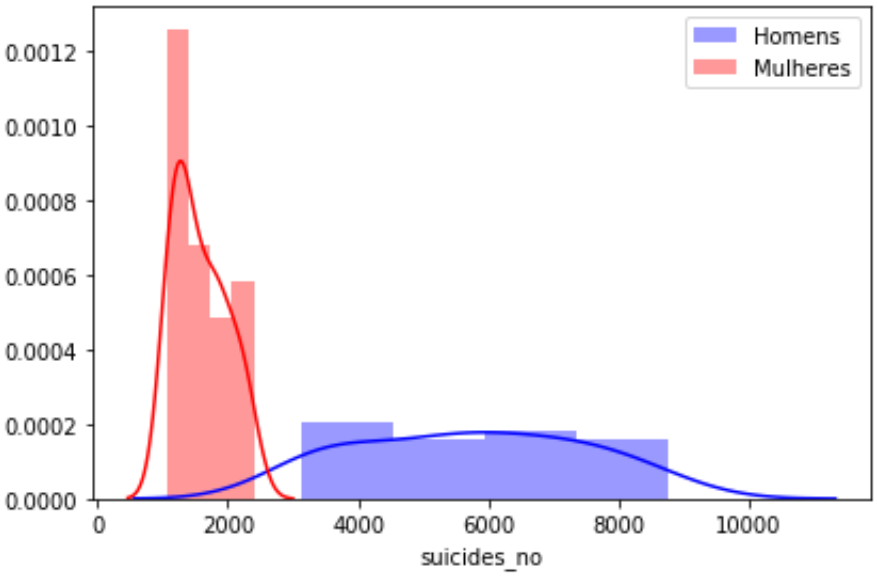

Figura 9. Número de suicídios homem vs mulher

Neste gráfico, podemos observar o número de variáveis ótimas obtidas pelo algoritmo de RFECV, tendo o melhor desempenho com 4 atributos.

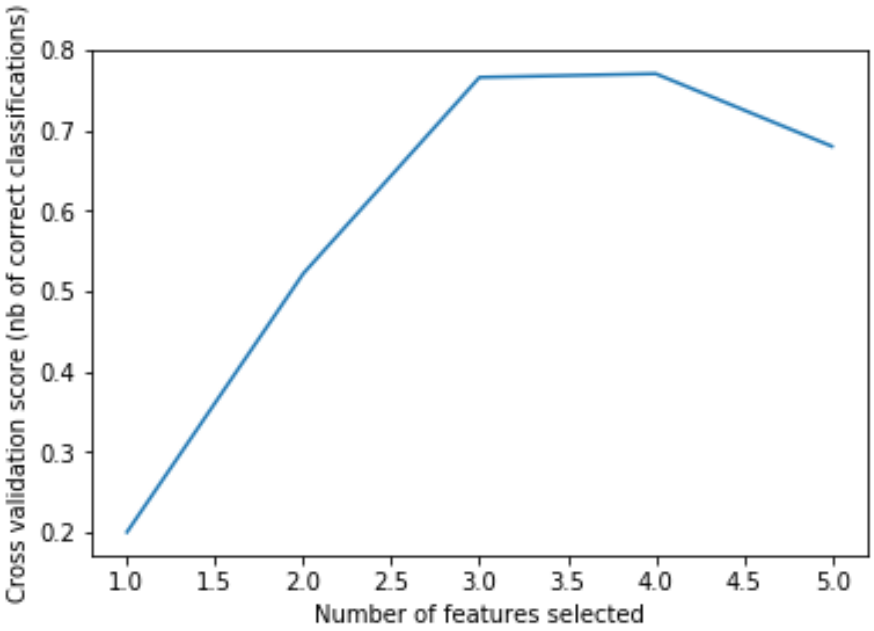

Figura 10. Desempenho do algoritmo vs Número de atributos

Com o *GridSearchCV* o desempenho obtido com os parâmetros ajustados da árvore de decisão no conjunto de treinamento foi um R^2 de 0.970383, enquanto no conjunto de teste obteve-se um R^2 de 0.948667.

A imagem da Árvore final, foi gerada usando *export_graphviz*, onde pôde-se visualizar nos nós os atributos conforme os critérios estabelecidos comentados na seção Limpeza e preparação dos dados. A Árvore está em anexo ao projeto devido a dificuldade da sua representação juntamente ao texto, no arquivo "**tree.png**".

Utilizando o algoritmo de RFECV, os atributos escolhidos para a Árvore de Decisão de Regressão foram: *'sex', 'age','gdp_for_year', 'generation'*. O desempenho obtido deste modelo no conjunto de testes foi um R^2 de 0.948676, uma mudança bastante insignificante, visto que ocorreu apenas a partir da quinta casa decimal.

Em contrapartida, utilizando o algoritmo *SelectKBest* com a estatística F como parâmetro de seleção, foram obtidas as seguintes variáveis: *'sex', 'age', 'gdp_for_year', 'gdp_per_capita'*. O desempenho obtido com a utilização dessas variáveis foi de R^2 = 0.949214, uma mudança mais significativa do que a anterior, visto que dessa vez, ela ocorreu na terceira casa decimal.

5. Discussão

Diferentemente do que era imaginado, o *PIB per capita* não está inversamente ligado à taxa de suicídios, mas diretamente relacionado. Como é possível perceber, conforme uma taxa cresce, a outra cresce de forma quase que proporcional (dada a regressão). Isso tem causa em diversos fatores, como o aumento populacional, mudanças nas condições de trabalho, crises econômicas e distribuição de renda que, infelizmente, são dados que não estão contidos no conjunto analisado, sendo o *PIB per capita* o que mais se aproxima.

Infelizmente, como foi apontado, a base de dados não é completa o suficiente para o tipo de análise que buscava-se desenvolver, o que restringe o modelo ao relacionamento do *PIB per capita* com a taxa de suicídios, sem levar em consideração as condições sociais das pessoas, IDH do local (estados, cidades ou ainda bairros), condições de trabalho, desigualdade social, dentre outras [21] [11]. O *PIB per capita* pode ser sim um bom indicador, mas ainda assim é muito abrangente tratando-se de contexto social, uma vez que é apenas um valor médio, sem possibilidade de análise de desvio padrão, variância, etc.

Um dos fatores importantes que foi apontado é o fato de os homens estarem muito mais propensos a cometer suicídio, o que pode estar relacionado ao preconceito no momento de buscar ajuda emocional e, consequentemente, com a não realização de tratamentos psicológicos ou psiquiátricos [6]. Novamente, infelizmente, trata-se apenas de especulação, uma vez que os dados trabalhados são insuficientes para realizar este tipo de análise, que pode ser feita em estudos futuros com bases mais completas.

Por fim, outro fator que pode ser analisado é o preocupante crescimento na taxa de suicídio entre crianças e adolescentes, que pode estar ocorrendo por conta do aumento da pressão para atingir metas para as quais ainda não estão preparadas. Os pais, por sua vez podem, por falta de atenção ou de conhecimento a respeito do tema, estar ignorando os sinais que as crianças dão e não procurar o tratamento adequado. Juntando estes fatores com o crescimento populacional, tem-se uma taxa efetiva maior no aumento de suicídios.

6. Conclusão

Este estudo fornece um modelo de dados que identifica possíveis pessoas com o risco de suicídio, no qual, pode ser replicado em estudos futuros e utilizado para pesquisas, organizações ou intervenções terapêuticas que tem como objetivo administrar algum tipo de ação preventiva, visto que é um tema complexo de ser abordado, pois, além de a quantidade de dados disponíveis sobre suicídios ou lesões autoprovocadas serem incompletos, vários fatores influenciam nessa decisão como fatores sociais, econômicos, biológicos e temporais.

É possível concluir, através das técnicas aplicadas para essa base de dados, que, claramente, a taxa de suicídios no Brasil está aumentando em todas as faixas etárias, embora por volta dos anos 2000 tenha ocorrido um declínio nessa curva. Conclui-se também que, por algum motivo, os homens tem uma maior tendência a cometer suicídio que mulheres, além de verificar que o PIB per capita, ao contrário das expectativas inciais, influenciou mais na taxa de suicídio que as gerações analisadas, (*Silent*, X, G.I, Y, *Boomies*, Z e *Millenials*), aumentando de forma diretamente proporcional ao aumento do *PIB per capita*.

Referências

[1] Lloyd G Allred and Gary E Kelly. Supervised learning techniques for backpropagation networks. In *1990 IJCNN International Joint Conference on Neural Networks*, pages 721–728. IEEE, 1990.

[2] F Amirmoradi Kh Khosravi Z Godarzi A.M Memari, T Ramim. Causes of suicide in married women. *Journal of hayat*, 12(1):47–53, 2006.

[3] Enrique Baca-García, M Mercedes Perez-Rodriguez, Ignacio Basurte-Villamor, Jeronimo Saiz-Ruiz, José M Leiva-Murillo, Mario de Prado-Cumplido, Ricardo Santiago-Mozos, Antonio Artés-Rodríguez, Jose De Leon, et al. Using data mining to explore complex clinical decisions: a study of hospitalization after a suicide attempt. *Journal of Clinical Psychiatry*, 67(7):1124–1132, 2006.

[4] Courtney Bagge and Augustine Osman. The suicide probability scale: Norms and factor structure. *Psychological reports*, 83(2):637–638, 1998.

[5] Leo Breiman, Jerome H Friedman, Richard A Olshen, and Charles J Stone. Classification and regression trees. belmont, ca: Wadsworth. *International Group*, page 432, 1984.

[6] Escola Nacional de Saúde Pública Sérgio Arouca. Pesquisa revela: homens não procuram serviços de saúde. `http://www.ensp.fiocruz.br/portal-ensp/informe/site/materia/detalhe/22251`, 2010. Acessed: 2019-05-09.

[7] Simone C. Garcia. O uso de árvores de decisão na descoberta de conhecimento na área da saúde. *LUME - Repositório Digital da UFRGS*, 2003.

[8] Rise B Goldstein, Donald W Black, Amelia Nasrallah, and George Winokur. The prediction of suicide: Sensitivity, specificity, and predictive value of a multivariate model applied to suicide among 1906 patients with affective disorders. *Archives of general psychiatry*, 48(5):418–422, 1991.

[9] Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. Gene selection for cancer classification using support vector machines. *Machine learning*, 46(1-3):389–422, 2002.

[10] Jiawei Han, Jian Pei, and Micheline Kamber. *Data mining: concepts and techniques*. Elsevier, 2011.

[11] Daniel Kim. The associations between us state and local social spending, income inequality, and individual all-cause and cause-specific mortality: The national longitudinal mortality study. *Preventive Medicine*, 84:62 – 68, 2016.

[12] Sotiris B Kotsiantis, I Zaharakis, and P Pintelas. Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering*, 160:3–24, 2007.

[13] Teresa LaFromboise and Beth Howard-Pitney. The zuni life skills development curriculum: Description and evaluation of a suicide prevention program. *Journal of Counseling Psychology*, 42(4):479, 1995.

[14] NUMFOCUS. pandas: powerful python data analysis toolkit. `https://pandas.pydata.org/pandas-docs/stable/`, 2019. Acessed: 2019-5-5.

[15] OMS. Oms: suicídio é responsável por uma morte a cada 40 segundos no mundo. `https://nacoesunidas.org/oms-suicidio-e-responsavel-por-uma-morte-a-cada-40-segundos-no-mundo/`. Acessed: 2019-5-5.

[16] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[17] Scikit-learn. FRegression documentation. `https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.f_regression.html#sklearn.feature_selection.f_regression`. Acessed: 2019-05-03.

[18] Scikit-learn. GridSearchCV documentation. `https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html`. Accessed: 2019-05-01.

[19] Scikit-learn. SelectKBest documentation. `https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.SelectKBest.html`. Acessed: 2019-05-03.

[20] Juyoung Song, Tae Min Song, Dong-Chul Seo, and Jae Hyun Jin. Data mining of web-based documents on social networking sites that included suicide-related words among korean adolescents. *Journal of Adolescent Health*, 59(6):668–673, 2016.

[21] Steven Stack and Ira Wasserman. Economic strain and suicide risk: A qualitative analysis. *Suicide and Life-Threatening Behavior*, 37(1):103–112, 2007.