Projeto 2 - Ciência de Dados

Turma: 2 B
Integrantes:
- Fabrizio Stresser Milan
- Pedro de Souza Zequi
- Thiago Hampl de Pierri Rocha
- Tomás Fiorelli Barbosa

Introdução

O objetivo do Projeto 2 de Ciência de Dados consiste em uma análise exploratória de um *dataset*, com o intuito de aplicar modelos de predição e prever uma variável *target* (dependente) em função de *features* (independentes).

O grupo possuía interesse em trabalhar com algum tipo de dado governamental e/ou algo mais realista e concreto. Então, ficou-se por muito tempo procurando por um dataset através dos sites disponibilizados, como o Kaggle e o site do INEP. No final, foi decidido que o *dataset* sobre a causa de morte dos cidadãos nova-iorquinos seria o escolhido para a realização do projeto.

Com o *dataset* em mente, partiu-se da seguinte pergunta formulada: "a quantidade de mortes entre 2007 e 2014 deveria ser considerada alarmante no estado de Nova Iorque, tendo em conta a média de mortes de todos os anos?". Enfim, com tudo pronto para uso, foi possível iniciar o projeto, decidindo os modelos utilizados e extrair seus resultados.

Minerando Dados e Características do Dataset

Os dados utilizados neste projeto foram retirados do *NYC open data*, um portal de dados livres para o público coletados por diversas agências da cidade de Nova Iorque. No caso do nosso dataset, foi publicado pela *Department of Health and Mental Hygiene (DOHMH)*. Encontramos esse portal a partir do Kaggle, que, por possuir dados de diferentes segmentos, lugares ou agências, seria muito difícil procurar por um que se atentasse à nossa ideia. No fim, o processo foi bom, já que começamos num portal com um escopo maior, fomos a um mais especializado para, enfim, achar um conjunto de dados governamentais que nos interessava.

	Year	Leading Cause	Sex	Race Ethnicity	Deaths	Death Rate	Age Adjusted Death Rate
0	2010	Influenza (Flu) and Pneumonia (J09-J18)	F	Hispanic	228	18.7	23.1
1	2008	Accidents Except Drug Posioning (V01-X39, X43,...	F	Hispanic	68	5.8	6.6
2	2013	Accidents Except Drug Posioning (V01-X39, X43,...	M	White Non-Hispanic	271	20.1	17.9
3	2010	Cerebrovascular Disease (Stroke: I60-I69)	M	Hispanic	140	12.3	21.4
4	2009	Assault (Homicide: Y87.1, X85-Y09)	M	Black Non-Hispanic	255	30	30
...	...	...	...	...	...	...	...
1089	2012	Influenza (Flu) and Pneumonia (J09-J18)	F	Not Stated/Unknown	6	.	.
1090	2014	Accidents Except Drug Posioning (V01-X39, X43,...	F	White Non-Hispanic	169	11.9	7.4
1091	2009	Malignant Neoplasms (Cancer: C00-C97)	M	White Non-Hispanic	3236	240.5	205.6
1092	2009	Intentional Self-Harm (Suicide: X60-X84, Y87.0)	M	White Non-Hispanic	191	14.2	13
1093	2013	Essential Hypertension and Renal Diseases (I10...	M	Black Non-Hispanic	148	17.2	20.9

Figura 1 - Dataset inicial, extraído no Kaggle

Inicialmente, foi difícil entender ao certo o que era cada linha, pois membros do grupo entendiam que era cada morte de pacientes por alguma doença (o que não faria sentido pois havia uma feature de mortes). Após ler bastante na descrição do dataframe, ficou claro o que significava: cada linha possuia o número de óbitos de pacientes da mesma etnia, sexo, por uma doença em um ano. Por exemplo, na primeira linha dos dados brutos acima, o *dataframe* explica que houve 228 mortes de hispânicas por pneumonia ou gripe apenas no ano de 2010.

O primeiro passo da simplificação dos dados brutos foi retirar as *features Death Rate* e *Age Adjusted Death Rate* já que não explicam nada sobre o porquê das pessoas morrerem, mas sim dados a mais da quantidade de mortes. Retirar a coluna do ano também foi uma simplificação adequada para o projeto, já que na pergunta motivadora do projeto já está constatado que é referente aos anos de 2007 a 2014 e, portanto, se considerarmos os anos como variáveis explicativas da quantidade de mortes o projeto se complicará muito para uma primeira iteração. É plausível que o ano tenha relação com as mortes (talvez tenha ocorrido uma infestação de uma doença contagiosa em determinado ano), no entanto o projeto visa unicamente as características pessoais do paciente e a causa, que são atemporais, para possivelmente prever uma morte nos dias de hoje.

Posteriormente, para se adequar à pergunta de sim ou não, o próximo passo foi transformar a coluna *Deaths* em um valor binário (0 se as mortes estão abaixo da média calculada de 388.48 e 1 se estão acima). Dessa forma, tal coluna se tornará explicativa para, no exemplo da primeira linha, responder se existem quantidades de mortes preocupantes no caso das hispânicas com gripe ou pneumonia (nesse caso para a simplificação adotada nesse projeto, não é alarmante). Por fim, bastou substituir as causas de mortes pelas iniciais, para simplificar o dataframe.

<u>Siglas e o nome de suas doenças correspondentes:</u>
AEDP: *Accidents Except Drug Posioning* (V01-X39, X43, X45-X59, Y85-Y86)
AD: *All Other Causes*
AOC: *Alzheimer's Disease* (G30)
AAD: *Aortic Aneurysm and Dissection* (I71)
As: *Assault* (*Homicide*: Y87.1, X85-Y09)
At: *Atherosclerosis* (I70)
CD: *Cerebrovascular Disease* (*Stroke*: I60-I69)
CCPP: *Certain Conditions originating in the Perinatal Period* (P00-P96)
CLDC: *Chronic Liver Disease and Cirrhosis* (K70, K73)
CLRD: *Chronic Lower Respiratory Diseases* (J40-J47)
CMDCA: *Congenital Malformations, Deformations, and Chromosomal Abnormalities* (Q00-Q99)
DM: *Diabetes Mellitus* (E10-E14)
DH: *Diseases of Heart* (I00-I09, I11, I13, I20-I51)
EHRD: *Essential Hypertension and Renal Diseases* (I10, I12)
HIVD: *Human Immunodeficiency Virus Disease* (HIV: B20-B24)
IP: *Influenza (Flu) and Pneumonia* (J09-J18)
IBUN: *Insitu or Benign / Uncertain Neoplasms* (D00-D48)
ISH: *Intentional Self-Harm* (*Suicide*: X60-X84, Y87.0)

MN: *Malignant Neoplasms* (Cancer: C00-C97)
MBDAPOPSU: *Mental and Behavioral Disorders due to Accidental Poisoning and Other Psychoactive Substance Use* (F11-F16, F18-F19, X40-X42, X44)
MBDUA: *Mental and Behavioral Disorders due to Use of Alcohol* (F10)
NNSN: *Nephritis, Nephrotic Syndrome and Nephrisis* (N00-N07, N17-N19, N25-N27)
PD: *Parkinson's Disease* (G20)
S: *Septicemia* (A40-A41)
T: *Tuberculosis* (A16-A19)
VH: *Viral Hepatitis* (B15-B19)

Sexo:
F: Feminino
M: Masculino

Raça/Etnia:
- *Asian and Pacific Islander*
- *Black Non-Hispanic*
- *Hispanic*
- *Not Stated/Unknown*
- *White Non-Hispanic*
- *Other Race/Ethnicity*

	Leading Cause	Sex	Race Ethnicity	y
0	IP	F	Hispanic	0
1	AEDP	F	Hispanic	0
2	AEDP	M	White Non-Hispanic	0
3	CD	M	Hispanic	0
4	As	M	Black Non-Hispanic	0
...	...	...	...	...
1089	IP	F	Not Stated/Unknown	0
1090	AEDP	F	White Non-Hispanic	0
1091	MN	M	White Non-Hispanic	1
1092	ISH	M	White Non-Hispanic	0
1093	EHRD	M	Black Non-Hispanic	0

Figura 2 - Resultado final dos dados prontos para a análise

Para uma análise inicial dos dados, estudamos se a coluna *target* 'y' está equilibrada entre os valores 0 e 1.

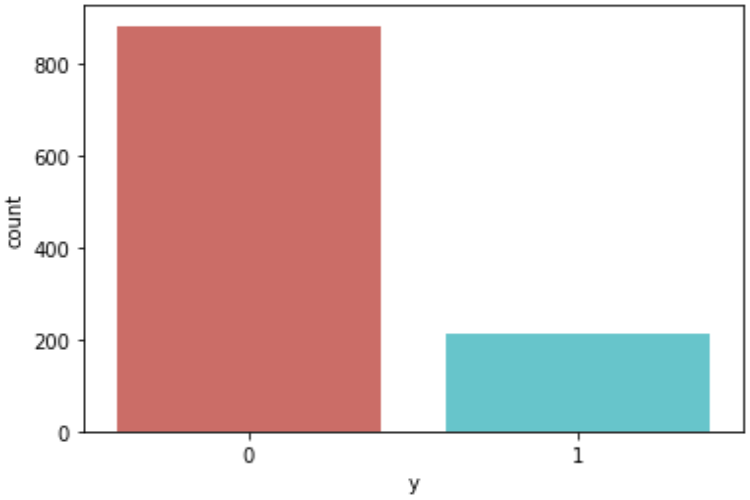

Figura 3 - Contabilização dos resultados em 'y'

A partir dos valores acima:
- 80,53% dos dados reconhecem que não há um estado alarmante de mortes.
- 19,47% dos dados reconhecem que há um estado alarmante de mortes.

Para um primeiro estudo das variáveis,cruzamos o target com cada feature. Como estamos cruzando variáveis quantitativas com qualitativas em todos os casos, estudamos a medida resumo de cada rótulo. Para a causa da morte:

Doença (Sigla)	**Média do 'y'**
AD	0.521
CD	0.056
CLRD	0.125
DM	0.043
DH	0.667
IP	0.167
MN	0.656

A partir dos valores acima, é possível perceber que dentre as 26 causas de mortes, apenas 7 possuem valores 1 em 'y'. Ou seja, 7 causas são consideradas problemáticas para o valor de 388 mortes. Dentre essas 7, 3 possuem média maior de 0,5 das linhas como valor 1, sendo elas:

- Outras doenças (0,521);
- Problema no coração (0,667);
- Tumor maligno (0,656).

Inicialmente é plausível, portanto, considerar essas 3 como as mais problemáticas causas de morte em Nova Iorque, no entanto não estamos considerando o sexo nem a etnia das pessoas.

Para o sexo:

Sexo	**Média do 'y'**
F (mulher)	0.202
M (homem)	0.187

Por ter poucos rótulos, a coluna de sexo talvez não tenha relação com a causa de morte. Pelas médias acima, ambos estão com valores baixos. Mas isso acontece exatamente porque o *dataframe* possui muito mais valores 0, dessa forma não é possível determinar se tal característica está relacionada com as mortes.

Para a etnia:

Etnia	**Média do 'y'**
Asian and Pacific Islander	0.186
Black Non-Hispanic	0.292
Hispanic	0.271
White Non-Hispanic	0.455

Nesse caso, os valores são mais dispersos, já que das 6 etnias, 4 possuem algum valor 1 na coluna 'y', sendo a maior média dos brancos não-hispânicos. As duas etnias que não possuem algum valor 1 são 'desconhecida' e 'outra etnia'. Dessa forma, é possível estimar que essas duas sem nenhum valor 0 tenham bastante correlação com o valor de 'y', com coeficientes negativos (já que se possuir tal etnia, o valor 'y' tende bem mais a 1 do que 0) e, talvez, em módulo os maiores.

Modelos de Predição

Modelo 1 - Regressão Logística

O primeiro modelo de predição escolhido pelo grupo foi o de Regressão Logística, um modelo estatístico utilizado para estimar a probabilidade de ocorrência de uma variável categórica dependente. Por meio do Machine Learning e com o uso do Python, a variável

dependente é definida como uma variável binária, no caso do projeto, essa variável foi definida na coluna *Deaths,* quando as mortes estão abaixo da média, é atribuído o valor 0 e, quando as mortes estão acima da média, é atribuído o valor 1. Para que o modelo seja implementado de forma eficiente, é necessário uma quantidade considerável de *samples*, filtrar a base de dados de forma que sejam utilizadas apenas variáveis significativas, que o modelo apresente pouca linearidade, que as variáveis independentes sejam independentes entre si. Para o tema escolhido pelo grupo, o modelo de regressão logística é plausível pois todas *features* são variáveis qualitativas, algo que tal modelo sabe como prever (diferente de uma regressão linear). O primeiro passo do modelo foi transformar as 3 colunas em valores dummies. Isso faz com que as 3 colunas qualitativas se transformem em várias colunas quantitativas de valor binário, que representa se a linha possui tal rótulo ou não. No caso do nosso dataframe, 4 colunas se transformaram em 35 (coluna 'y', 26 causas de morte, 2 de sexo e 6 etnias). Após essa adequação, o modelo estava pronto para ser aplicado.

	y	Leading Cause_AEDP	Leading Cause_AD	Leading Cause_AOC	Leading Cause_AAD	Leading Cause_As	Leading Cause_At	Leading Cause_CD	Leading Cause_CCPP	Leading Cause_CLDC	...	Leading Cause_T	Leading Cause_VH	Sex_F
0	0	0	0	0	0	0	0	0	0	0	...	0	0	1
1	0	1	0	0	0	0	0	0	0	0	...	0	0	1
2	0	1	0	0	0	0	0	0	0	0	...	0	0	0
3	0	0	0	0	0	0	0	1	0	0	...	0	0	0
4	0	0	0	0	0	1	0	0	0	0	...	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1089	0	0	0	0	0	0	0	0	0	0	...	0	0	1
1090	0	1	0	0	0	0	0	0	0	0	...	0	0	1
1091	1	0	0	0	0	0	0	0	0	0	...	0	0	0
1092	0	0	0	0	0	0	0	0	0	0	...	0	0	0
1093	0	0	0	0	0	0	0	0	0	0	...	0	0	0

1094 rows × 35 columns

Figura 4 - Exemplo de algumas variáveis em formato dummy

<u>Modelo 2 - Árvore de Decisão</u>

O segundo modelo de predição escolhido pelo grupo foi a Árvore de Decisão, uma ferramenta preditiva binária utilizada principalmente para mapear diversas possíveis consequências e a probabilidade de um certo evento ocorrer ou não, é um modelo baseado em resultados sequenciais que disputam entre si. Para o tema escolhido pelo grupo, é um ótimo modelo desde que busca por diversos resultados e consequências a fim de chegar a uma conclusão, onde é possível visualizar de forma clara quais caminhos que o programa seguiu para chegar a uma resposta para a pergunta proposta pelo grupo. Além disso, é um modelo de fácil visualização e entendimento por ser facilmente representado por gráficos ou imagens, com a ajuda de ferramentas de programação. Na parte gráfica, cada 'nó' ou 'folha' da árvore possui duas possíveis saídas, a qual é atribuído um determinado valor para cada amostra ou resultado. A árvore parte de uma variável inicial, uma raíz, em que são criadas as duas primeiras saídas e, o modelo segue essa sequência de eventos.

Processo e Estatísticas de Validação

Antes de implementar ambos os modelos, foi necessário separar o *dataframe* em treinamento e teste. A parte de treinamento será usada para construir os modelos e, a de teste, para validar o modelo. Dessa forma, se houver qualquer manipulação a ser feita, como equilibrar os valores 0 e 1 de 'y', devem ser feitos unicamente na parte de treinamento, para não descredibilizar as validações. No caso deste projeto, escolhemos 30% das linhas dos dados para ser o teste e 70%, treinamento.

Modelo 1 - Regressão Logística

Um possível problema de previsão para o nosso *dataframe* é a falta de balanceamento entre os dados 0 e 1, que podem causar uma tendência maior do modelo a classificar 'y' como 0. Para resolver este problema, importamos a função '*SMOTE*' da biblioteca *imblearn.over_sampling* que, com uma linha de código, cria linhas "artificiais" no banco de dados de treinamento que balanceiam o *dataframe*. Essa manipulação da função é baseada nas linhas existentes de códigos, portanto a acurácia não cairá tanto e o modelo, mesmo que acerte menos, será mais condizente com valores de mortes alarmantes (objetivo final do trabalho).

A partir da biblioteca '*statsmodels.api*', atribuímos um modelo de regressão logística tanto para as linhas não balanceadas como para as com '*SMOTE*'. A acurácia do primeiro modelo, comparando as features de teste e 'y' de teste com a predição sem '*SMOTE*', resultou em, aproximadamente, 0.9757 (97%).

A do segundo modelo, por sua vez, tem, 0.9574 (95%), aproximado. Assim como esperado, criar linhas artificiais diminui um pouco a acurácia, mas é um valor pequeno e, portanto, o grupo achou plausível manter o segundo modelo, pois possivelmente terá um valor melhor para os valores 1 de 'y'. Abaixo segue a matriz de confusão do modelo escolhido:

```
[254  13]
[ 1   61]
```

Dos valores que possuem 0 no 'y' de teste, o modelo foi capaz de acertar 254 e errar 13. Dos valores que possuem 1 no 'y' teste, o modelo errou apenas 1 vez. Para as mortes não alarmantes, acertou 95,13% e, para as alarmantes, 98,39% de acurácia.

Mesmo com resultados bons, o modelo considera colunas com valores 'P' muito altos, ou seja, que não possuem relação forte com a coluna 'y'. Segue abaixo lista completa:

Leading Cause_AEDP	0.00621
Leading Cause_AD	0.00000
Leading Cause_AOC	0.00149
Leading Cause_AAD	0.99926
Leading Cause_As	0.38922
Leading Cause_At	0.99926
Leading Cause_CD	0.97879
Leading Cause_CCPP	0.42070
Leading Cause_CLDC	0.01350
Leading Cause_CLRD	0.41436
Leading Cause_CMDCA	0.81927
Leading Cause_DM	0.07303
Leading Cause_DH	0.00003
Leading Cause_EHRD	0.00409
Leading Cause_HIVD	0.02447
Leading Cause_IP	0.00875
Leading Cause_IBUN	0.99929
Leading Cause_ISH	0.00036
Leading Cause_MN	0.00000
Leading Cause_MBDAPOPSU	0.00082
Leading Cause_MBDUA	0.99926
Leading Cause_NNSN	0.21615
Leading Cause_PD	0.99985
Leading Cause_S	0.17915
Leading Cause_T	0.99985
Leading Cause_VH	0.90065
Sex_F	0.42497
Sex_M	0.63748
Race Ethnicity_Asian and Pacific Islander	0.66830
Race Ethnicity_Black Non-Hispanic	0.00001
Race Ethnicity_Hispanic	0.56783
Race Ethnicity_Not Stated/Unknown	0.00000
Race Ethnicity_Other Race/ Ethnicity	0.00000
Race Ethnicity_White Non-Hispanic	0.00000

Figura 5 - Variáveis dummies *e seus respectivos valores de 'P'*

O método seguinte foi bem metódico: um por um, retirar a coluna com maior valor 'P' do modelo, rodar a regressão logística novamente, analisar os novos valores 'P', retirar o maior, até que todas as colunas tenham este valor menor do que 0,1 ('α' plausível para um bom modelo). Após 22 iterações retirando colunas com valor 'P' alto, as colunas que sobraram descreviam muito bem o funcionamento da variável 'y'.

Results: Logit

Model:	Logit	Pseudo R-squared:	0.682
Dependent Variable:	y	AIC:	567.0760
Date:	2020-11-24 14:54	BIC:	633.5468
No. Observations:	1228	Log-Likelihood:	-270.54
Df Model:	12	LL-Null:	-851.18
Df Residuals:	1215	LLR p-value:	3.7364e-241
Converged:	0.0000	Scale:	1.0000

	Coef.	Std.Err.	z	P>|z|	[0.025	0.975]
Leading Cause_AD	2.8464	0.3558	8.0005	0.0000	2.1491	3.5437
Leading Cause_AOC	-4.5656	2.1827	-2.0917	0.0365	-8.8436	-0.2875
Leading Cause_CLDC	-3.8722	1.8912	-2.0476	0.0406	-7.5788	-0.1656
Leading Cause_DH	8.4459	3.8587	2.1888	0.0286	0.8830	16.0089
Leading Cause_HIVD	-4.0114	2.0961	-1.9137	0.0557	-8.1198	0.0969
Leading Cause_ISH	-4.8494	2.2575	-2.1481	0.0317	-9.2740	-0.4247
Leading Cause_MN	7.8405	3.0476	2.5727	0.0101	1.8673	13.8137
Leading Cause_MBDAPOPSU	-4.4373	2.0483	-2.1663	0.0303	-8.4520	-0.4226
Leading Cause_S	-2.4787	1.3856	-1.7889	0.0736	-5.1945	0.2370
Race Ethnicity_Black Non-Hispanic	-1.1102	0.2365	-4.6949	0.0000	-1.5737	-0.6467
Race Ethnicity_Not Stated/Unknown	-14.7474	6.0930	-2.4204	0.0155	-26.6893	-2.8054
Race Ethnicity_Other Race/ Ethnicity	-14.2998	5.5675	-2.5684	0.0102	-25.2119	-3.3877
Race Ethnicity_White Non-Hispanic	0.7450	0.1783	4.1783	0.0000	0.3955	1.0944

Figura 6 - Resultados obtidos no Python sobre o modelo de Regressão Logística

```
[262   5]
[ 11  51]
```

O modelo acima possui 13 colunas de informação, ao invés dos 35 anteriores, e uma acurácia de 0.9514 (95%), aproximando o valor. É minimamente menor do que o anterior mas muito mais simples de se construir. A matriz de confusão realça a melhoria em prever valores 0 em y (98,13%, um acréscimo de 3 pontos percentuais), no entanto a predição de valores 1 caiu (de 98,39% para 82,25%).

A partir da comparação acima, deve ser necessário balancear se o modelo quer ter como preferência a acurácia dos valores 1 em y ou a facilidade de minerar dados. Para o acúmulo inicial de dados, seria mais barato estudar apenas as 13 colunas de informação que possuem valor baixo.

Na análise exploratória inicial, as maiores médias de y estavam em AD (*All Other Causes* || "Todas as outras causas"), DH (*Diseases of Heart* || "Problemas do coração"), MN (*Malignant Neoplasms* || "Tumor maligno"). Nos resultados acima, são as 3 aparecem, logo possuem valor 'P' baixo. Além disso, possuem os maiores coeficientes do modelo de forma positiva. Todas as outras causas são pouco explicativas de forma geral, mas possuem um coeficiente significativamente menor que as outras duas causas. Logo, como primeira conclusão final da regressão logística, Problemas do Coração e Tumor Maligno devem ser considerados causadores de quantidades de mortes alarmantes para o governo de Nova Iorque.

Além disso, assim como previsto na análise exploratória, etnias desconhecidas e outras etnias fazem parte das características que se relacionam com o valor de 'y' com

valores de coeficientes negativos e, em módulo, altos (14.747 e 14.2998, respectivamente). Mesmo que não sejam etnias específicas, isso sugere que as possíveis minorias em Nova Iorque não correspondem à grande quantidade de mortes (etnias que não aparecem nos dados talvez tenham poucos pacientes e, portanto, uma teoria plausível é que se trata de minorias muito específicas, como os amarelos).

Modelo 2 - Árvore de Decisão

A biblioteca importada para tal modelo foi '*sklearn.tree*', utilizando a função '*DecisionTreeClassifier*'. A primeira iteração desse modelo foi usando os dados não balanceados pelo SMOTE. Este resultou numa acurácia de aproximadamente 0.9878 (98%). Para um primeiro modelo, foi um resultado extremamente gratificante.

Mesmo com um valor bom, utilizamos o SMOTE para o modelo a ser estudado, a fim de melhorar e balancear as probabilidades. Nesse caso, a acurácia subiu para 0.9909 (99%), aproximadamente. Diferente da regressão logística, o processo de balanceamento rendeu previsões melhores usando as mesmas linhas de dataframe. Isto deixa claro, como uma primeira iteração, a qualidade da árvore de decisão neste projeto. Abaixo segue a matriz de confusão deste modelo balanceado:

```
[266   1]
[  2  60]
```

Este resultado tem a predição de valores 0 melhor que qualquer modelo de regressão logístico feito (99,62% de acertos), e uma previsão bem mais satisfatória dos valores 1 do que o modelo com as colunas relevantes da regressão logística (96,77% de acurácia). Em geral, apresenta resultados excelentes.

Além disso, a árvore de decisão tem outro ponto positivo, já que é possível visualizar exatamente o caminho traçado pelo algoritmo para chegar em resultados tão bons. è difícil de colocar no relatório, já que possui muitas informações, mas é possível dar um zoom nos primeiros a fim de exemplificar o caminho.

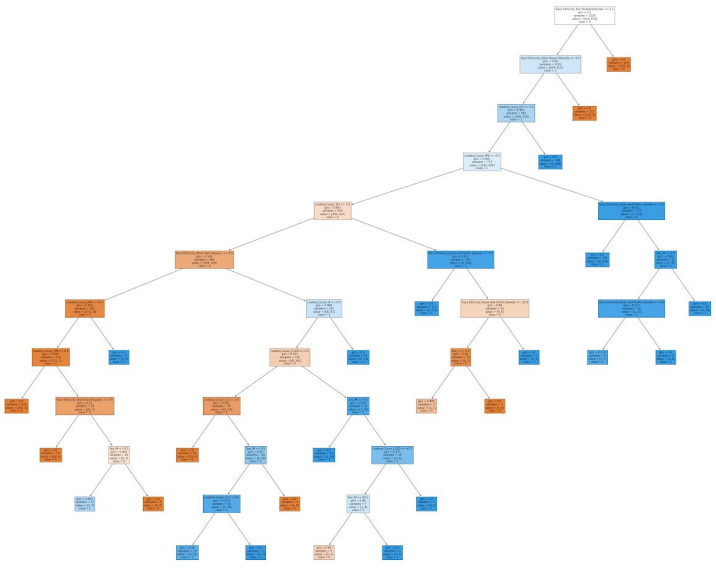

Figura 7 - Resultado completo do modelo da Árvore de Regressão (pouco visível)

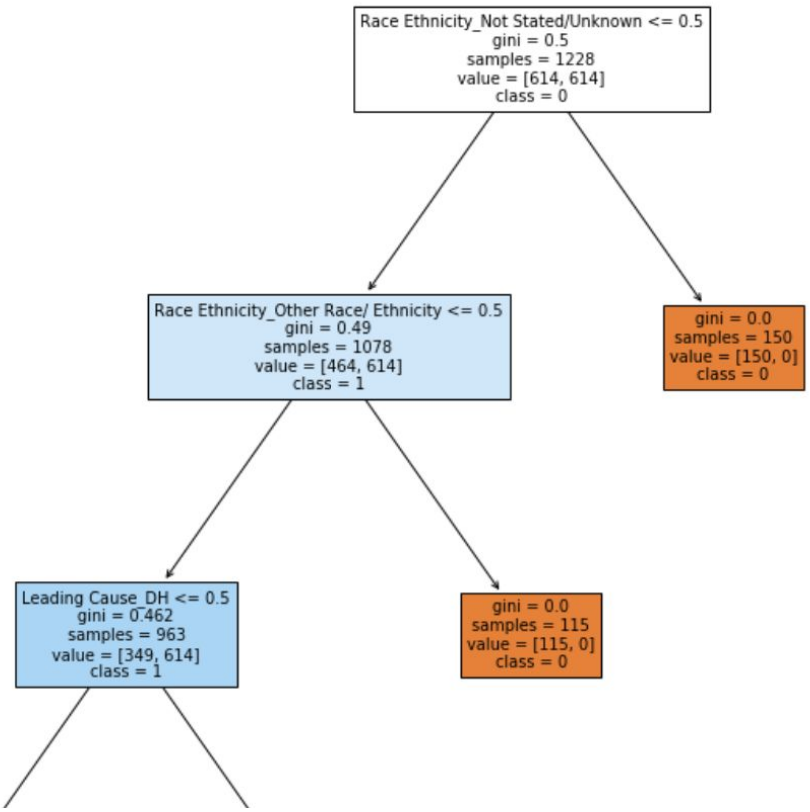

Figura 8 - Resultado completo do modelo da Árvore de Regressão (primeiros galhos)

Conclusão

Dessa forma, como conclusão final dos modelos, é plausível garantir que a árvore de decisão prevê melhor que a regressão logística, mas requer as 35 colunas de informação já que não foi retirado as colunas com pouca relação com y. O grupo acredita que, se houver recursos suficientes para minerar todos os dados, esse é o modelo principal. Se for necessário baratear e simplificar a pesquisa inicial, o modelo de regressão logística final cumpre bem seu papel, já que ainda possui uma acurácia geral de 95,14%.

Diante do exposto, pode-se dizer que o projeto de análise exploratória e desenvolvimento de modelos de predição, utilizando um *dataset* ‘cru’, foi devidamente executado, apresentando resultados plausíveis e satisfatórios para o pouco tempo de execução.

Referências Bibliográficas

[1] Dataset
- Kaggle. **New York City Leading Causes of Death**.
Disponível em:
<https://www.kaggle.com/new-york-city/ny-new-york-city-leading-causes-of-death>
Acesso em: 06/11/2020.

[2] Estrutura do Projeto
- Ensinando Máquinas. **Modelos Preditivos de Notas de Redação do ENEM 2015**.
Disponível em:
<https://ensinandomaquinasblog.wordpress.com/2017/12/15/modelos-preditivos-de-notas-de-redacao-do-enem-2015/>
Acesso em: 06/11/2020.

[3] Regressão Logística
- Wikipédia. **Regressão Logística**.
Disponível em:
<https://en.wikipedia.org/wiki/Logistic_regression>
Acesso em: 10/11/2020.

[4] Visualização da Árvore de Decisão
- Mljar. **Visualize a Decision Tree in 4 Ways with Scikit-Learn and Python**.
Disponível em:
<https://mljar.com/blog/visualize-decision-tree/>
Acesso em: 19/11/2020.