

2019-05-27-dt

May 29, 2019

#

Equidade em Aprendizado de Máquina

## 0.1 Renan Del Buono Brotto

## 0.2 Introdução

As técnicas de Aprendizado de Máquina estão em rápida ascensão em diversas áreas, como processamento de linguagem natural e reconhecimento de fala [?], processamento de imagem [?], problemas de classificação [?], reconhecimento de padrões [?] entre outras.

O objetivo central do Aprendizado de Máquina é projetar modelos computacionais capazes de aprender uma determinada tarefa a partir de dados, sem que exista uma programação específica para a tarefa em questão [?]. Para tanto, apresentamos um conjunto de dados disponíveis, denominado conjunto de treinamento, a fim de que o modelo capture as informações relevantes para o problema investigado [?].

Além de ter um bom desempenho, segundo uma métrica adequada, frente aos dados de treinamento, o modelo ajustado deve ser capaz de apresentar um bom desempenho frente a dados desconhecidos, pertencentes a um conjunto normalmente denominado conjunto de teste. Este segundo requisito configura a capacidade de generalização do modelo [?], [?].

Contudo, durante o processo de aprendizagem, o modelo pode capturar informações irrelevantes, ou mesmo indesejadas, para a tarefa considerada, configurando uma situação de sobreajuste [?]. Neste tipo de situação, temos um bom desempenho frente ao conjunto de treinamento, mas o modelo se comporta mal quando avaliado frente a novos dados [?]. Dentre as informações indesejadas aprendidas, podemos citar ruído presente sobre os dados [?] ou mesmo algum tipo de viés apresentado nos dados de treinamento.

O viés contido nos dados de treinamento desempenha um papel fundamental quando aplicamos os modelos de Aprendizado de Máquina em tarefas com impacto social direto, tais como concessão de crédito [?], aprovação em universidades [?] ou a concessão de liberdade condicional [?]. Neste tipo de problema as técnicas clássicas de aprendizado incorporam as tendências observadas nos dados, o que pode dar origem a atributos discriminativos, como por exemplo o gênero em um problema de concessão de crédito ou a etnia em um problema de concessão de liberdade condicional. Uma vez aprendida esta característica, o modelo ajustado passa a produzir resultados que podem contribuir ainda mais com a disparidade do problema [?].

## 0.3 Objetivo

Nosso objetivo neste trabalho é aplicar as técnicas de ICA [?] sobre os atributos do nosso problema de modo a promover a equidade na classificação. Para tanto, vamos investigar a descorrelação

linear, tendo como base o trabalho [?], e a descorrelação não-linear, que é uma condição mais próxima da independência.

Na sequência, descreveremos o *dataset* que usaremos.

## 0.4 Descrição do Dataset

Neste trabalho, usaremos o Adult Data Set (Disponível em: <http://archive.ics.uci.edu/ml/datasets/adult> e de domínio público). Este conjunto de dados é formado por 14 atributos, tanto categóricos quanto atributos numéricos, anonimizado. Dentre estes atributos temos idade, tipo de trabalho, gênero, etnia, estado civil, horas trabalhadas por semana entre outros. Para cada indivíduo do conjunto de dados, temos um rótulo, indicando uma renda anual superior (igual) a \$50 k ou inferior a este limiar.

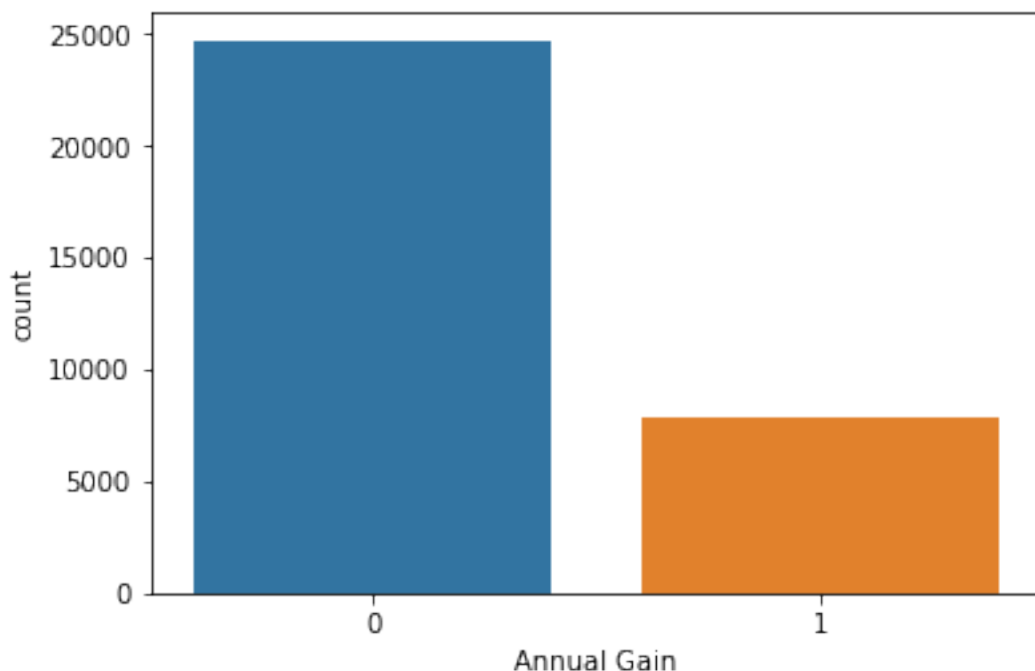
Como muitos dos atributos são categóricos, vamos convertê-los para atributos numéricos.

Out[1]: <IPython.core.display.HTML object>

Para melhor entendermos a inequidade presente nos dados, vamos comparar, na sequência, o percentual de indivíduos com alta renda na população completa, na população masculina e na população feminina.

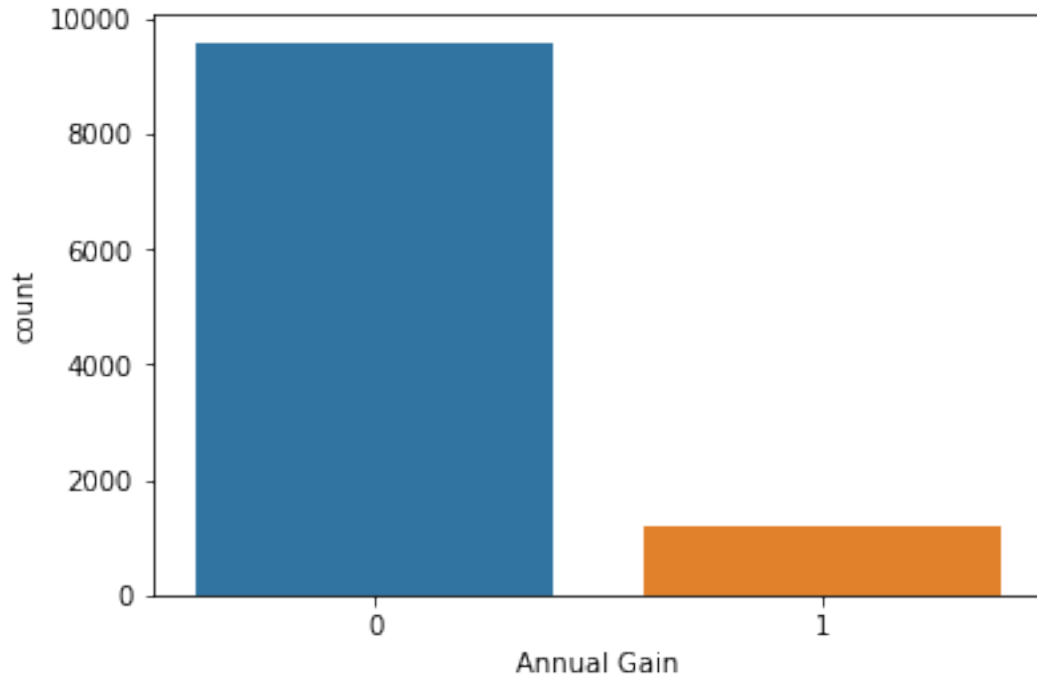
### 0.4.1 População Completa

Vamos agora, visualizar um pouco os dados. Inicialmente, vamos analisar a proporção de indivíduos com alta renda ( $\geq 50K$ ) na população completa.



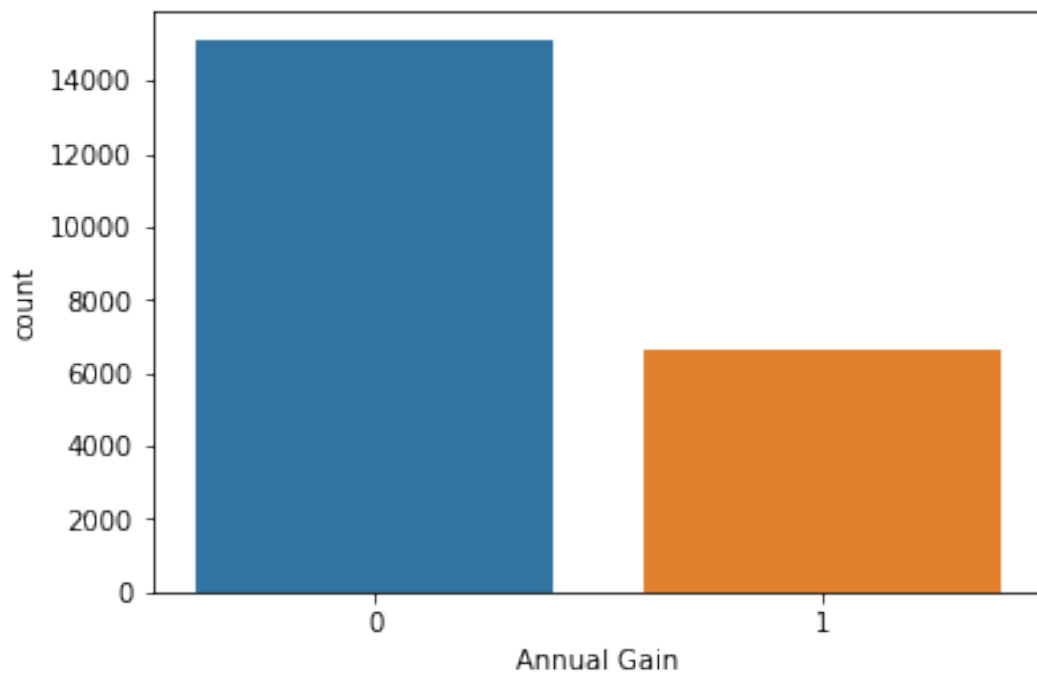
Vamos agora repetir a mesma análise entre as populações masculina e feminina. ### Análise das Populações Masculina e Feminina

Out[10]: <matplotlib.axes.\_subplots.AxesSubplot at 0xd93a7f0>



Percebemos do resultado acima que o número de indivíduos na população feminina na classe "Alta Renda" é inferior ao percentual da população como um todo.

Out[12]: <matplotlib.axes.\_subplots.AxesSubplot at 0xb905780>



```

-----

TypeError                                Traceback (most recent call last)

<ipython-input-31-3c04269735b4> in <module>
    12
    13 cm = sns.light_palette("green", as_cmap=True)
--> 14 sns.heatmap(percentual_formatado, annot=True)

~\Anaconda3\lib\site-packages\seaborn\matrix.py in heatmap(data, vmin, vmax, cmap, center, robust, annot, fmt,
515     plotter = _HeatMapper(data, vmin, vmax, cmap, center, robust, annot, fmt,
516                          annot_kws, cbar, cbar_kws, xticklabels,
--> 517                          yticklabels, mask)
518
519     # Add the pcolormesh kwargs here

~\Anaconda3\lib\site-packages\seaborn\matrix.py in __init__(self, data, vmin, vmax, cmap, center, robust, annot, fmt,
165     # Determine good default values for the colormap
166     self._determine_cmap_params(plot_data, vmin, vmax,
--> 167                                cmap, center, robust)
168
169     # Sort out the annotations

~\Anaconda3\lib\site-packages\seaborn\matrix.py in _determine_cmap_params(self, plot_data, vmin, vmax, cmap, center, robust):
202     cmap, center, robust):
203     """Use some heuristics to set good defaults for colorbar and range."""
--> 204     calc_data = plot_data.data[~np.isnan(plot_data.data)]
205     if vmin is None:
206         vmin = np.percentile(calc_data, 2) if robust else calc_data.min()

TypeError: ufunc 'isnan' not supported for the input types, and the inputs could not be

```

O que notamos do resultado acima é uma inequidade entre indivíduos rotulados como "Alta Renda" nas populações masculina e feminina. Treinar um classificador da maneira tradicional, sem nenhum tratamento sobre os dados, fará com que o modelo de Aprendizado de Máquina incorpore as tendências observadas nos dados, o que pode contribuir com a acentuação da inequidade observada.

Na sequência do texto, apresentamos a nossa proposta para promover a equidade na classificação.

# 1 Metodologia

Para a tarefa de classificação, empregaremos um classificador ajustado segundo o paradigma de regressão logística. Para os nossos propósitos aqui, adotaremos o atributo "Gênero (*Sex*)" como atributo discriminatório. Ao todo, empregaremos 3 classificadores:

- 1) Classificador Tradicional: neste caso, faremos o treinamento sem nenhum tipo de ajuste sobre os dados.
- 2) Classificador com Descorrelação Linear: neste segundo caso, buscaremos descorrelacionar linearmente o atributo discriminatório das classes do problema ("Fairness Constraints: Mechanisms for Fair Classification", Zafar, *et al.*, 2017).
- 3) Classificador com Descorrelação Não Linear: neste último caso, buscaremos a independência entre o atributo Gênero e as classe do problema, por meio da descorrelação não-linear.

Apresentamos com maior detalhe o nosso *Workflow* no diagrama a seguir:

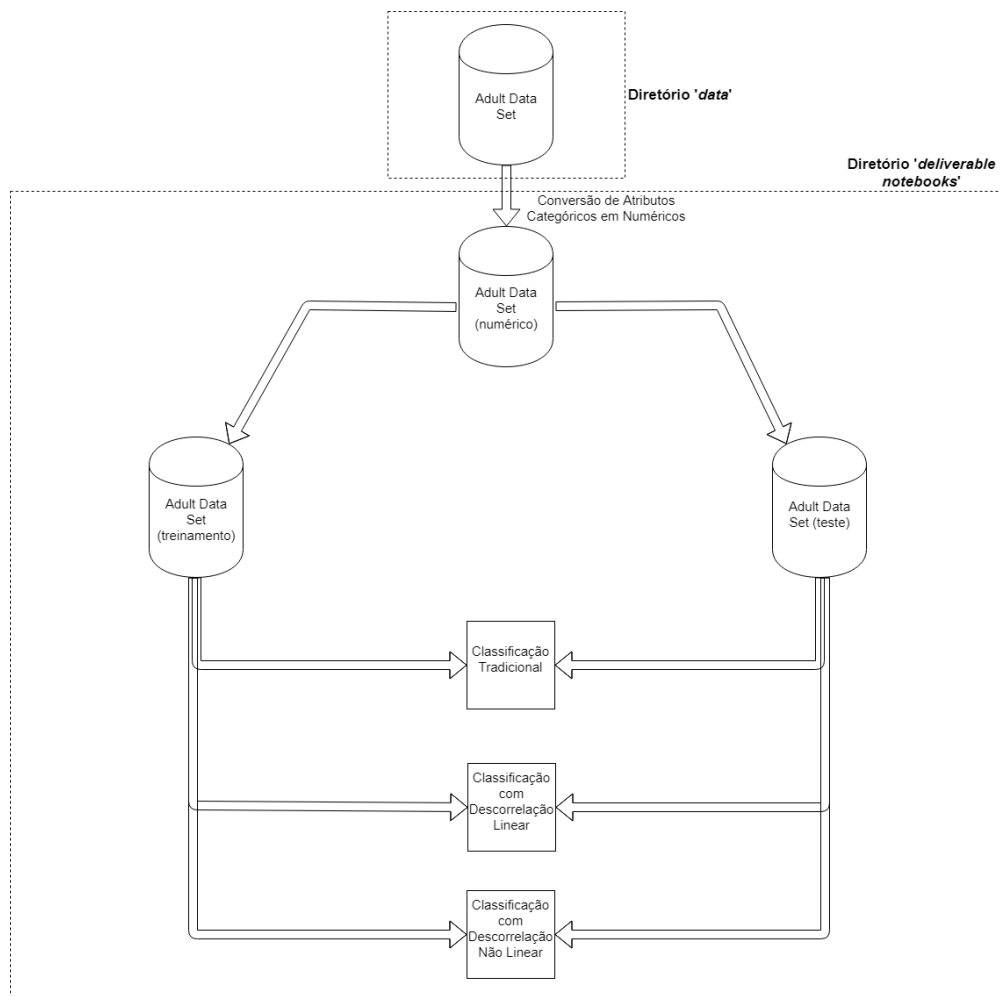
## 2 Workflow

### 2.1 Resultados

Nesta seção, apresentamos nossos resultados na tarefa de classificação em termos de acurácia e equilíbrio entre os indivíduos classificados como "Alta Renda" nas populações masculina e feminina.

#### Classificador 1

Árvore de Decisão 1 Out [17]:



Workflow



mos a influência do atributo discriminatório sobre as classes do problema, eliminamos parte da informação de que o classificador necessita.

Esta perda de informação é muito mais acentuada no Classificador 3. Como aplicamos a decorrelação não-linear, eliminamos de modo mais acentuado a influência do atributo discriminatório sobre as classes do problema, quando comparado com a redução provida pela decorrelação linear.

Na sequência, vamos avaliar como cada um dos classificadores atua na promoção da equidade. Para isso, avaliamos o percentual de indivíduos classificados como "Alta Renda" em cada uma das subpopulações de interesse.

	Pop. Masculina [%]	Pop. Feminina [%]
Classificador 1	34.234234234234236	6.132268632268632
Árvore de Decisão 1	21.263308763308764	3.8185913185913183
Classificador 2	16.625716625716624	3.501228501228501
Árvore de Decisão 2	20.915233415233413	4.217854217854218
Classificador 3	14.680589680589682	3.1326781326781328
Árvore de Decisão 3	20.904995904995904	4.2588042588042585

Da tabela acima, notamos que o Classificador 2, apesar de ter uma acurácia inferior à do Classificador 1, teve um pequeno avanço em direção à equidade, uma vez que houve uma redução no percentual de indivíduos da população masculina classificados como "Alta Renda" e um acréscimo desta mesma categoria na população feminina.

Já no caso do Classificador 3, notamos um acréscimo de indivíduos pertencentes à classe "Alta Renda" nas duas subpopulações, o que não caminha em direção à equidade.

Para o cenário que estudamos, ambas as estratégias para mitigar o efeito do atributo discriminatório levaram a uma queda da acurácia, como esperado. Contudo, a decorrelação linear permitiu avanços em direção à equidade, algo que não observamos na versão não linear (Classificador 3).

## 2.2 Conclusões

Neste trabalho avaliamos como algumas das técnicas de ICA, *i.e.*, decorrelação linear e decorrelação não-linear, podem ser usadas para promover a equidade na tarefa de classificação. Para o caso particular que estudamos, ambas as técnicas levaram a uma redução de acurácia nos classificadores. Contudo, a decorrelação linear levou a uma pequena aproximação entre as populações masculina e feminina, em termos de indivíduos classificados como "Alta Renda". Não observamos a mesma aproximação para quando empregamos a decorrelação não-linear.