Aprendizado de Máquinas – Avaliação 1

Professor: Rodrigo Targino

Aluno: Caio Lins

27 de março de 2022

1 Apresentação do Dataset

Como banco de dados para ser utilizado no projeto final da disciplina, escolhemos o *"Census Income Data Set"*[1]. O dataset foi extraído da base de dados do censo populacional americano de 1994. Cada instância representa um indivíduo, que possui algumas caracteríscas sociais e econômicas disponíveis, correspondentes às colunas da tabela. Os dados foram extraídos já com um problema de classificação binária em mente: a coluna relativa à renda anual do indivíduo só informa se ele ganha mais ou menos que $50 000 por ano. Nosso objetivo será ajustar modelos de classificação a esses dados, de modo a classificar um dado indivíduo com relação ao seu salário, baseando-nos nas outras variáveis disponíveis. Nessa primeira parte, faremos uma análise exploratória dos dados, discutindo suas principais características.

1.1 Descrição das variáveis

Para ler e manipular os dados, utilizamos a versão 1.2.4. da biblioteca *Pandas* [2], da linguagem de programção *Python* [3], versão 3.8.10. No repositório da UCI, os dados se encontram previamente divididos em um conjunto para treino e outro para teste, na proporção 2:1. Nós juntamos as duas tabelas para obter um data frame com todas as instâncias disponíveis. No total, são 48 840 entradas com 15 atributos distintos, listados a seguir:

- `Age`. Variável numérica que assume valores inteiros. Corresponde à idade do indivíduo.
- `workclass`. Variável categórica que assume os seguintes valores: `Private`, `Self-emp-not-inc`, `Self-em-inc`, `Local-gov`, `State-gov`, `Federal-gov`, `Without-pay`, `Never-worked`. Corresponde a uma classificação do emprego do indivíduo, com relação ao tipo de vínculo empregatício.
- `fnlwgt`. Variável numérica que assume valores inteiros. É um peso calculado pelo *United States Census Bureau* que indica quantas pessoas aquele indivíduo representa na população . Ele é necessário para realizar inferências sobre a população toda do país, devido às estratégias de amostragem utilizadas para decidir quem será entrevistado no censo.

- `education`. Variável categórica que assume os seguintes valores: `Preschool`, `1st-4th`, `5th-6th`, `7th-8th`, `9th`, `10th`, `11th`, `12th`, `HS-grad`, `Some-college`, `Assoc-voc`, `Assoc-acdm`, `Bachelors`, `Masters`, `Prof-school`, `Doctorate`. Corresponde ao grau máximo de educação obtido pelo indivíduo.

- `education_num`. Variável numérica que assume valores inteiros entre 1 e 16. Corresponde a uma codificação numérica da educação do indivíduo, na ordem apresentada na descrição da variável `education`. Não está claro o que alguns valores que `education` assume significam, como, por exemplo, `Assoc-voc` ou `Assoc-acdm`. Com isso, não temos total certeza que a variável `education` é ordenável. Entretanto, vamos supor que ela é, justamente por causa dessa ordenação que foi disponibilizada junto com o dataset.

- `marital_status`. Variável categórica que assume os seguintes valores: `Married-civ-spouse`, `Divorced`, `Never-married`, `Separated`, `Widowed`, `Married-spouse-absent`, `Married-AF-spouse`. Corresponde à situação conjugal do indivíduo.

- `occupation`. Variável categórica que assume os seguintes valores: `Tech-support`, `Craft-repair`, `Other-service`, `Sales`, `Exec-managerial`, `Prof-specialty`, `Handlers-cleaners`, `Machine-op-inspct`, `Adm-clerical`, `Farming-fishing`, `Transport-moving`, `Priv-house-serv`, `Protective-serv`, `Armed-Forces`. Corresponde a uma classificação do emprego do indivíduo com relação ao tipo de atividade exercida.

- `relationship`. Variável categórica que assume os seguintes valores: `Wife`, `Own-child`, `Husband`, `Not-in-family`, `Other-relative`, `Unmarried`. Corresponde a uma classificação das relações familiares mais próximas do inidivíduo.

- `race`. Variável categórica que assume os seguintes valores: `White`, `Black`, `Asian-Pac-Islander`, `Amer-Indian-Eskimo`, `Other`. Corresponde à raça do indivíduo.

- `sex`. Variável categórica que assume os seguintes valores: `Male`, `Female`. Corresponde ao sexo do indivíduo.

- `capital_gain`. Variável numérica que assume valores inteiros. Não há uma descrição precisa do que essa variável significa na página do dataset. Entretanto, pelo nome dela pode-se inferir que se trata de algum ganho monetário não diretamente relacionado ao salário anual do indivíduo.

- `capital_loss`. Variável numérica que assume valores inteiros. Não há uma descrição precisa do que essa variável significa na página do dataset. Entretanto, pelo nome dela pode-se inferir que se trata de uma perda monetária não diretamente relacionada ao salário anual do indivíduo.

- `hours_per_week`. Variável numérica que assume valores inteiros. Corresponde ao número de horas semanais que o indivíduo passa trabalhando.

- `native_country`. Variável categórica que corresponde ao país onde o indivíduo nasceu. Há muitos valores possíveis para essa variável para citá-los todos.

	age	education_num	capital_gain	capital_loss
mean	38.643857	10.078092	1079.067301	87.505897
std	13.710652	2.570954	7452.168393	403.012415
min	17.000000	1.000000	0.000000	0.000000
25%	28.000000	9.000000	0.000000	0.000000
50%	37.000000	10.000000	0.000000	0.000000
75%	48.000000	12.000000	0.000000	0.000000
max	90.000000	16.000000	99999.000000	4356.000000

	hours_per_week	target
mean	40.422400	0.239292
std	12.391697	0.426655
min	1.000000	0.000000
25%	40.000000	0.000000
50%	40.000000	0.000000
75%	45.000000	0.000000
max	99.000000	1.000000

Tabela 1: Estatísticas descritivas para as variáveis numéricas do dataset.

- `target`. Variável categórica que corresponde à renda anual do indivíuo. Ela assume os valores `<=50K` e `>50K`.

Infelizmente, a descrição dos dados fornecida na página da UCI não é muito detalhada, e a base de dados original do censo de 1994 não está mais disponível no website do *United States Census Bureau*. Com isso, não está muito claro o que alguns valores assumidos por variáveis categóricas significam. Isso pode dificultar a interpretação das análises futuras.

Algumas estatísticas descritivas para as variáveis numéricas são apresentadas na Tabela 1.1. É interessante observar a probabilidade de um indivíduo ganhar mais de $50 000 por ano, sem nenhuma informação adicional: 23.92 %. Também vemos que os indivíduos têm, em média, 38.6 anos de idade e trabalham 40 horas por semana, o que está dentro do que se espera para a população economicamente ativa de um país.

1.2 Tratamento realizado

Além de juntar os datasets disponibilizados no repositório da UCI, foi necessário formatar as entradas da tabela que são strings, para retirar um espaço em branco que estava no início de todas elas. Também mudamos a variável `target` para que assumisse os valores 1, se o indivíduo ganha mais de $50 000 por ano, e 0 se não ganha e transformamos as variáveis categóricas não ordenáveis em indicadoras.

Tomamos a decisão de remover a coluna `fnlwgt`, pois o peso atribuído a cada indivíduo só é útil quando o objetivo é fazer inferências para toda a população americana. Como estamos mais

Atributo	Número de entradas faltantes
`workclass`	2799
`occupation`	2809
`native_country`	857

Tabela 2: Número de dados faltantes.

interessados em aplicar modelos preditivos apenas na amostra incompleta presente no nosso dataset, essa variável não é relevante, pois sabemos que ela não influencia a variável `target`.

É claro, mas vale ressaltar, que quando formos ajustar modelos aos nossos dados, não vamos utilizar, simultaneamente, as informações das colunas `education` e `education_num`, por serem, do ponto de vista informacional, idênticas.

Com relação a dados faltantes, apenas três colunas apresentam essa característica, como podemos ver na Tabela 1.2. Como a quantidade de dados faltantes é pequena em relação ao total de instâncias no conjunto de dados, descartar os indivíduos sem informação completa é uma estratégia viável para conseguir aplicar modelos de classificação. Entretanto, isso pode não ser necessário, e essa decisão não será tomada neste ponto da análise.

2 Estudo das relações entre as variáveis

Vamos começar analisando as correlações entre as variáveis numéricas disponíveis, apresentadas na Figura 1.

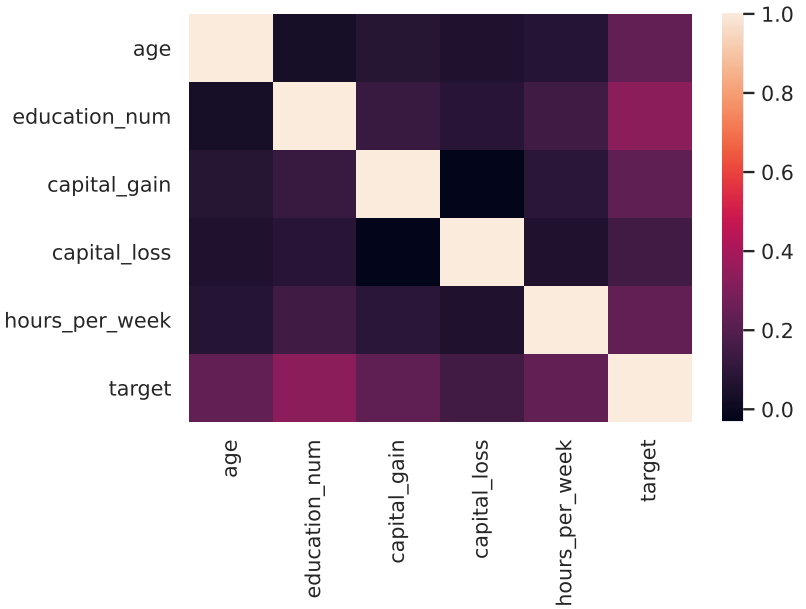

Figura 1: Correlações entre as variáveis numéricas do dataset.

Vemos que, no geral, as elas não são muito correlacionadas, especialmente as variáveis preditoras entre si. Os maiores coeficientes obtidos foram para as correlações que envolviam a variável inde-

pendente, sendo que o maior deles foi 0.332624, com a variável `education_num`. Isso indica que, em geral, indivíduos com maior escolaridade têm um salário mais elevado. É interessante observar que a variável `capital_loss` possui uma correlação *positiva* com `target`, o que é não é intuitivo, considerando seu nome. Isso nos faz questionar qual o real significado desas variável.

Também fizemos um plot contendo as correlações entre todas as variáveis (incluindo indicadoras) e o `target`, apresentado na Figura 2. Como são muitas categorias, selecionamos apenas aquelas que apresentaram um valor absoluto de correlação amostral maior que 0.2. É possível observar,

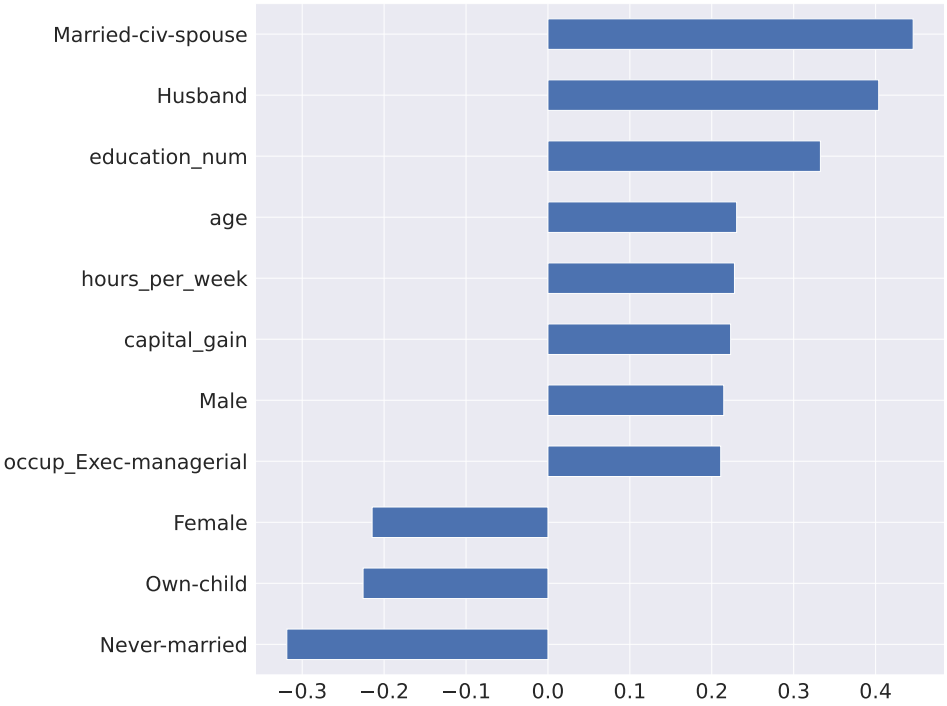

Figura 2: Variáveis que possuem correlação com a variável target com módulo superior a 0.2.

surpreendentemente, que as variáveis com maior correlação, tanto positiva quanto negativa, são relacionadas ao estado conjugal dos indivíduos. Vemos que estar em um casamento com um civil está correlacionado positivamente com ganhar mais de $50 000 por ano, enquanto nunca ter se casado está correlacionado negativamente.

É perceptível, também, que a característica `Male` possui correlação positiva com target, enquanto `Female` apresenta correlação negativa, indicando que, no geral, homens possuem salários melhores que mulheres. Também podemos ver que trabalhar mais durante a semana, ser mais velho e ter uma educação melhor estão correlacionados positivamente com possuir um salário acima de $50 000.

Com intuito de explorar mais essas relações, produzimos a Figura 4. Vemos que, à medida que o grau máximo de educação obtido aumenta, também aumenta a proporção de indivíduos que ganha mais de $50 000 por ano. Também percebemos que, no geral, poucas pessoas trabalham mais que 60 horas por semana, principalmente os mais velhos.

Por fim, também verificamos a distribuição da idade dos indivíduos, visualizando-a em conjunto com a variável `target` na Figura 3. Podemos ver que as maiores proporções de indivíduos que ganham mais de $50 000 são observadas entre 35 e 50 anos, aproximadamente.

3 Próximos passos

Na segunda parte desse trabalho, vamos ajustar modelos de classificação ao conjunto de dados escolhido, com o objetivo de prever o valor de `target` para um dado indivíduo. Para tanto, teremos que lidar de alguma forma com os dados faltantes, seja excluindo-os, ou aplicando técnicas estatísticas para preenchê-los artificialmente respeitando as características dos dados. Esperamos conseguir interpretar os resultados obtidos, de modo a identificar quais variáveis são mais relevantes para determinar a renda de um indivíduo.

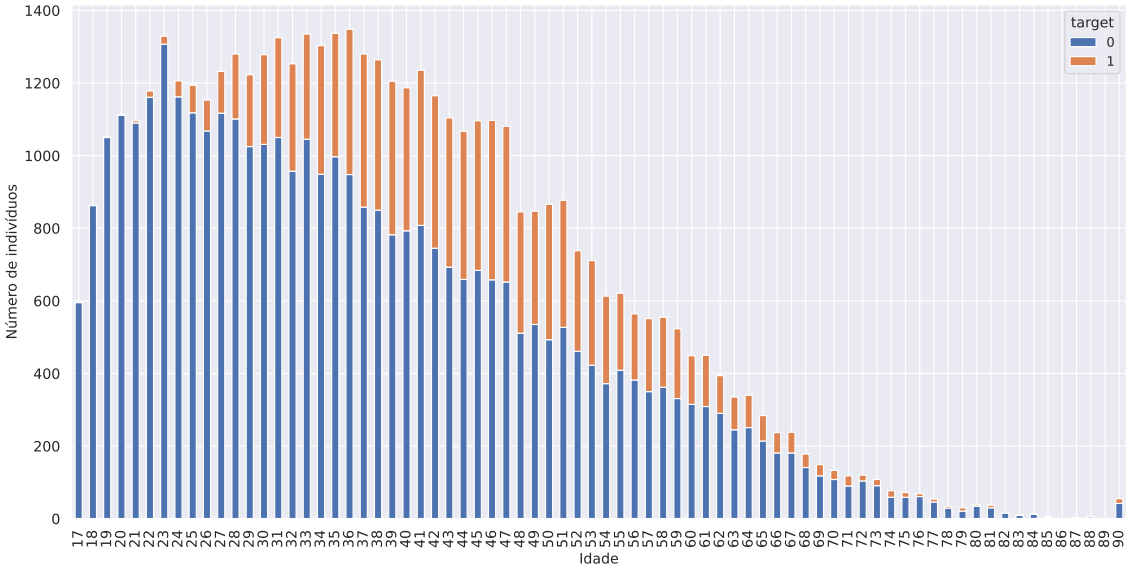

Figura 3: Distribuição da idade dos indivíduos, discriminada por target.

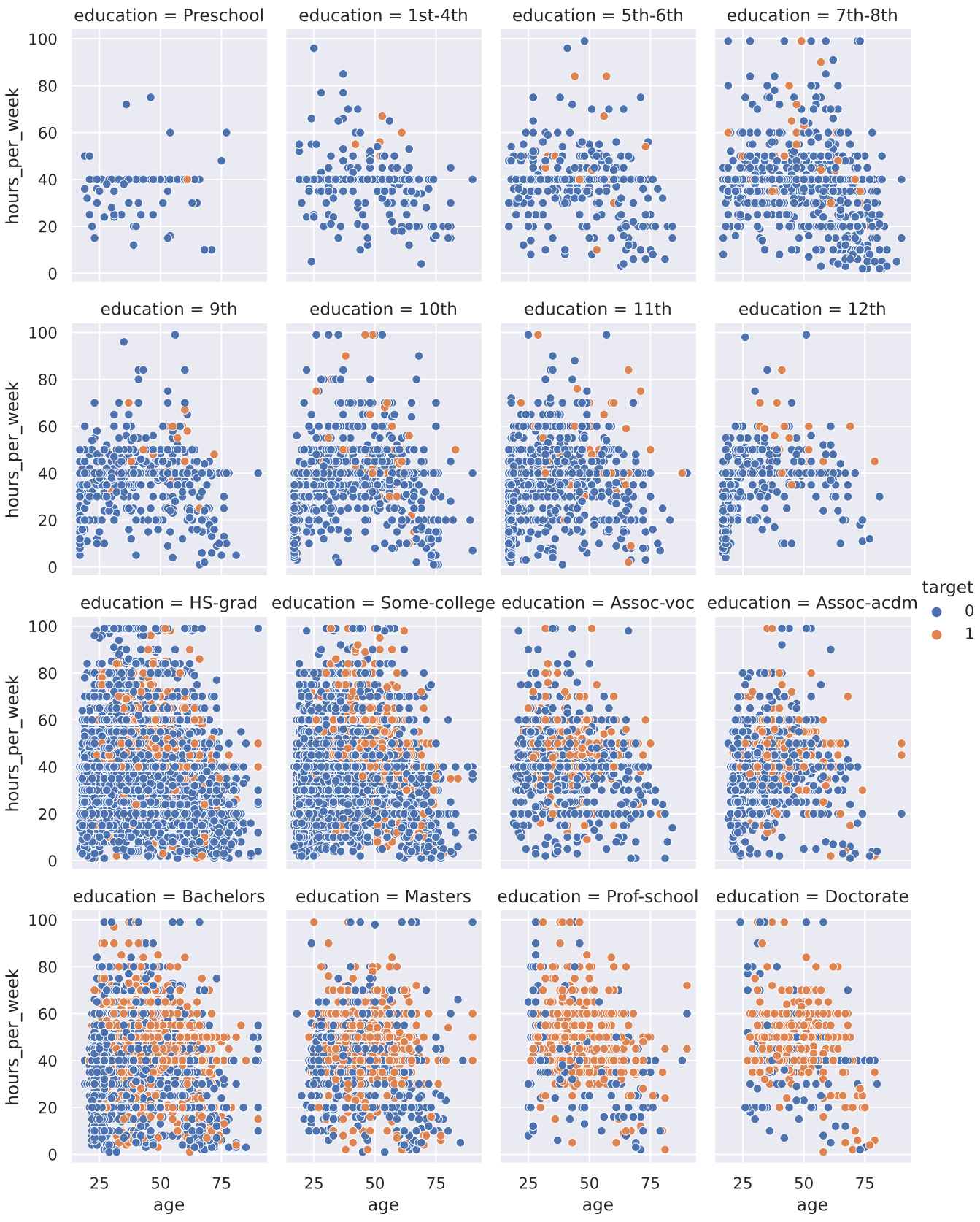

Figura 4: Horas trabalhadas por semana e idade, observadas em diferentes níveis de formação educacional.

Referências

[1] Dheeru Dua e Casey Graff. *UCI Machine Learning Repository – Census Income Data Set*. 2017. URL: `https://archive.ics.uci.edu/ml/datasets/Census+Income`.

[2] Wes McKinney. «Data Structures for Statistical Computing in Python». Em: *Proceedings of the 9th Python in Science Conference*. Ed. por Stéfan van der Walt e Jarrod Millman. 2010, pp. 51–56.

[3] Guido Van Rossum e Fred L. Drake. *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace, 2009. ISBN: 1441412697.