Aplicação do algoritmo de Árvore de Decisão para classificação de dados do uso de energia elétrica nos estados brasileiros

Danyelle da Silva Oliveira Angelo[1]

[1]Faculdade de Computação – Universidade Federal de Uberlândia (UFU)

danyelle.angelo@ufu.br

Resumo. *Neste trabalho aplicamos técnicas de machine learning (classificação), em um conjunto de dados que detalha a cadeia de energia elétrica do Brasil, com ênfase no consumo de energia elétrica por estado. O algoritmo escolhido para realizar a classificação desses dados foi o de árvore de decisão, apesar do alto grau de assertividade desse algoritmo, nossos resultados em termos de acurácia, não ultrapassaram o valor de 22%, mostrando que o modelo construído não é a melhor para escolha para o problema delineado.*

1. Introdução

Dados abertos é um termo amplamente usado na sociedade, e está relacionado a disponização de dados e informações de caráter governamental e público na internet. Mas não basta tornar os dados disponíveis, é preciso que estes sejam processáveis e sigam um padrão.

De acordo com o portal kit.dados.gov.br [gov.br], disponibilizar os dados governamentais de maneira padronizada permite:

- Economia de recursos: financeiro e de tempo, uma vez que o cidadão pode acessar as informações diretamente pela web, não há necessidade de protocolar uma ação junto à uma entidade.
- Redução de trabalho duplicado: os dados estando disponíveis, evita que outras equipes/entidades ou organizações, façam novamente a coleta de dados já existentes. Nessas situações os interessados podem reusar toda a base, ou até mesmo cruzar os dados já existentes, com novos dados;
- Conjuntos de dados podem ser combinados à conjuntos de dados de outras organizações, ampliando as possibilidades de trabalho sobre estes;
- Geração de empregos: a economia tende a ser estimulada, a medida que agentes econômicos utilizem dados abertos para a criação de novos processos de negócio e na otimização dos já existentes.

Pode-se dizer assim, que a disponibilização e padronização desses dados, permitem a potencialização do valor dos conjuntos em si. Permitindo que diferentes setores da sociedade participem do desenvolvimento de políticas, análises, pesquisas, aplicações privadas ou públicas e processos de negócio.

O site https://dados.gov.br/ disponibiliza mais de 10.000 conjuntos de dados para pesquisa, de 218 instituições, esses dados são apresentados em diferentes formatos. É

possível filtrar os conjuntos de dados disponíveis por organização, temas, etiquetas, licenças, ou formatos. Para a realização deste trabalho escolhemos trabalhar com um conjunto de dados que detalha o consumo de energia elétrica no país, nosso objetivo é classificar a unidade federativa (UF) de uma amostra com base nos valores de consumo, tipo de consumo, e região.

2. Conjunto de dados

O conjunto de dados escolhido foi produzido no ano de 2018 e foi construído pela *Empresa de Pesquisa Energética* (EPE). Conhecemos essa instituição através do portal dados.gov.br, entretanto, apesar de haver uma página dedicada a EPE no portal citado, a mesma se encontra vazia (https://dados.gov.br/organization/empresa-de-pesquisa-energetica-epe). Para ter acesso aos dados relativo ao consumo de energia elétrica no país é preciso acessar diretamente o site da EPE (https://www.epe.gov.br/pt/publicacoes-dados-abertos/dados-abertos/dados-do-anuario-estatistico-de-energia-eletrica). O conjunto disponibilizado no site da organização dá origem ao *Anuário Estatístico de Energia Elétrica*, que tem por objetivo publicar informações do mercado de eletricidade no Brasil, sendo um excelente exemplo dos benefícios da disponibilização dos dados abertos, como dito anteriormente.

2.1. Preparação dos dados

Nosso conjunto de dados, possuí 483.759 observações, realizamos uma série de operações sobre o mesmo antes de aplicar o modelo de classificação escolhido. Para realizar estas operações fizemos o uso do pacote *tidyverse*.

O *tidyverse* agrupa uma série de ferramentas pertinentes ao processo de aprendizagem de máquina. Entre os pacotes suportados pelo tidyverse, usados neste trabalho temos: *readr, dplyr* e *ggplot2*, sendo os 2 primeiros usados para leitura e manipulação de dados (respectivamente) e o segundo para a geração de gráficos (mostramos o uso desses pacotes na figura abaixo).

```
data <- read_csv2("epe.csv", col_select = selectedColumns)
data <- clean_names(data)
data <- as.data.frame(data)
data <- as.data.frame(unclass(data), stringsAsFactors = TRUE)

ggplot(treino) + geom_point(aes(consumo,sistema, color=setor_economico_n1))
avg_e_consumption <- treino |> group_by(uf, setor_economico_n1, sistema) |>
                          summarise(
                          `Gasto Médio` = mean(consumo,  na.rm = TRUE),
                          `Gasto Mediano` = median(consumo,  na.rm = TRUE),
                          `Menor Gasto` = min(consumo,  na.rm = TRUE),
                          `Maior Gasto` = max(consumo,  na.rm = TRUE),
                          `Quantidade de dados analisados` = n()
                      ) |>
                      arrange(desc(`Gasto Médio`))
```

Figure 1. Carregando dados, gerando gráfico e resumindo dados com tidyverse

Iniciamos o carregamento dos nossos dados (previamente baixados) através da função *read_csv2*, além do nome do arquivo que contém os dados, passamos também

uma lista com as colunas que desejávamos carregar em nosso ambiente, excluindo assim atributos que não tinham variabilidade ou não apresentavam relevância para o nosso problema.

Após formatar o conjunto de dados a ser analisado, utilizamos o pacote *ggplot2* e posteriormente a função *summarise* para entender a influência de cada atributo na base de dados. Construímos um gráfico de pontos, que relaciona o consumo de energia elétrica à uma região do país, categorizamos cada um desses pontos no gráfico através de cores, onde cada cor representa um setor econômico. Com a visualização do gráfico concluímos que as regiões com maior consumo de energia são a Sudeste e a Centro-Oeste, sendo que o tipo de consumo predominante é o Industrial.

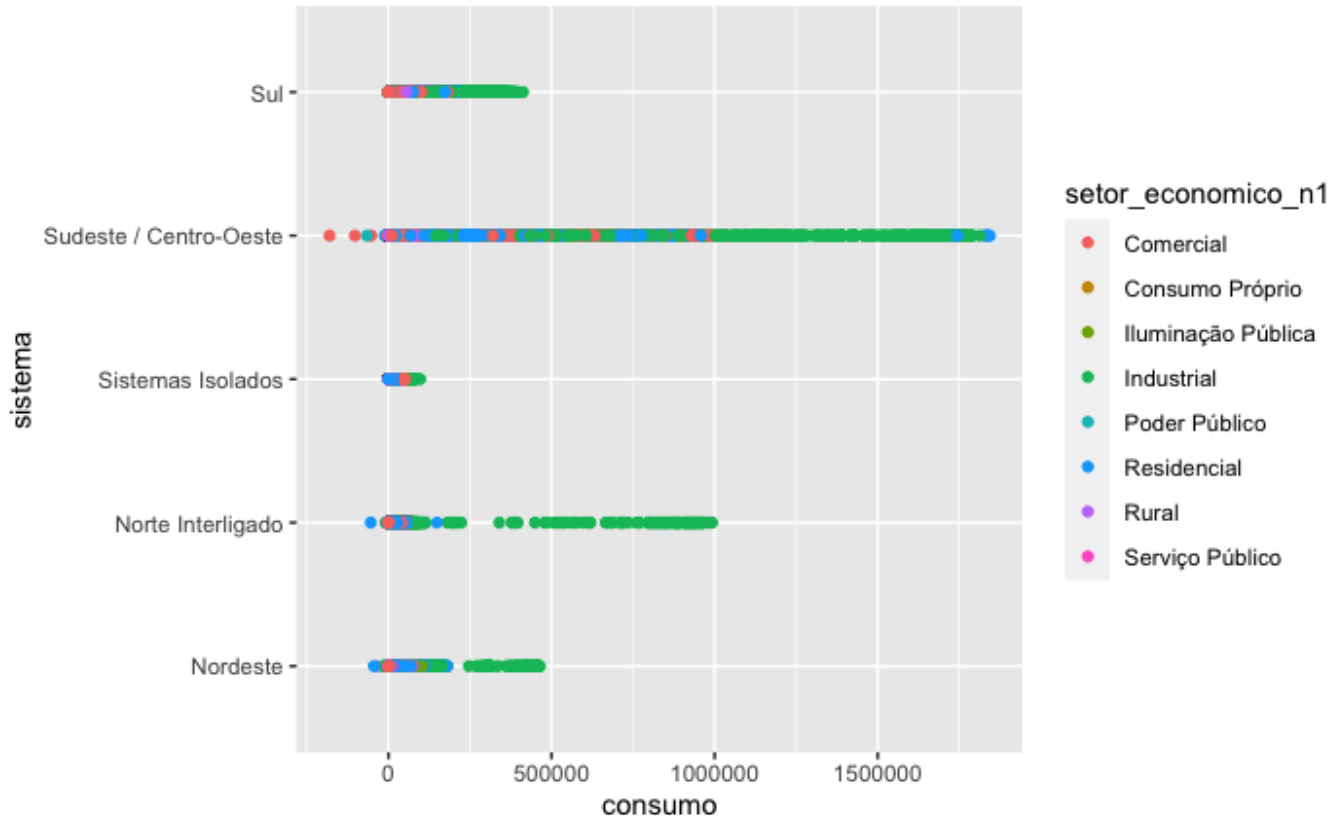

Figure 2. Tipo e valor de consumo energético por região

As conclusões acima, foram validadas através da sumarização dos dados da base de treino. Para realizar esse resumo dos dados agrupamos as amostras por estado (uf), tipo de consumo (setor_economico_n1) e região (sistema), após o agrupamento destes dados calculamos a mediana e média dos gatos, computamos também o maior e menor valor de gasto energético do agrupamento. Por fim, ordenamos estes dados em ordem decrescente de gasto médio. Os 10 agrupamentos com maior gasto energético (considerando tipo de gasto, e estado) são mostradas na Figura 3, e assim como inferimos do gráfico mostrado anteriormente, as regiões com maior consumo são Sudeste e Centro-Oeste.

uf <fctr>	setor_economico_n1 <fctr>	sistema <fctr>	Gasto Mediano <dbl>	Gasto Médio <dbl>	Menor Gasto <dbl>	Maior Gasto <dbl>
SP	Industrial	Sudeste / Centro-Oeste	47282.4530	3.653968e+05	-117.327	1813039.889
SP	Residencial	Sudeste / Centro-Oeste	170505.5000	2.524481e+05	-1453.000	1843483.000
MG	Industrial	Sudeste / Centro-Oeste	25260.5000	2.241523e+05	21.350	1617430.000
SP	Iluminação Pública	Sudeste / Centro-Oeste	123977.3695	1.268322e+05	-4516.761	282046.053
PA	Industrial	Norte Interligado	15814.5000	1.085064e+05	1.418	992216.000
PR	Industrial	Sul	55316.5000	9.833294e+04	1.000	367078.119
RJ	Iluminação Pública	Sudeste / Centro-Oeste	104811.9120	8.618226e+04	7.000	177647.000
BA	Iluminação Pública	Nordeste	87269.0000	8.603921e+04	60320.000	104493.222
RJ	Residencial	Sudeste / Centro-Oeste	58785.0000	8.295290e+04	267.069	357911.000
SC	Industrial	Sul	21698.2530	7.997583e+04	9.072	412381.114

1-10 of 273 rows | 1-7 of 8 columns Previous 1 2 3 4 5 6 ... 28 Next

Figure 3. Resumo dos dados

3. Aprendizagem de máquina

O problema proposto neste trabalho consiste em determinar o estado federativo (UF) relativo à uma amostra, dado uma região (sistema), Setor Econômico - N1, Setor Econômico - N2 e Taxa de Consumo elétrico. Sendo que os atributos Setor Econômico - N1 e Setor Econômico - N2 descrevem a classificação dos consumidores de energia elétrica de modo geral (o segundo é uma especificação do primeiro), são exemplos de classificação de consumo para os atributos citados acima:

- N1: Comercial, Consumo Próprio, Industrial, Poder Público, Rural e Serviço Público e etc;
- N2: Adm. Condominial, Iluminação, Uso Comum, Serviços de Comunicações e Telecomunicações, Templos Religiosos, Agroindústria, Escola Agrotécnica e etc.

Para realizar a classificação das nossas amostras, utilizamos o algoritmo de Árvore de decisão. Nesse algoritmo cada nó interno representa um teste de valor de um dos atributos das observações, e os nós folhas representam o resultado final da decisão, ou seja a classificação da amostra.

Com o uso do pacote *tidymodels* modelamos nossos dados, para aplicar o algoritmo de aprendizagem de máquina.

1. Separamos nossos dados de treino e de teste (80% para treino e 20% para teste);
2. Criamos uma especificação (receita) para o nosso modelo;
3. Definimos um modelo, no caso a escolha foi a árvore de decisão;
4. Criamos um fluxo de trabalho: esse processo é a junção do modelo definido com a receita;
5. E por fim executamos e medimos a qualidade do nosso classificador.

A imagem abaixo, traz o trecho do código em R, responsável pelos passos citados:

```
split_data <- initial_split(data, prop = 0.8)
treino <- training(split_data)
teste <- testing(split_data)

preparation_tree <- recipe(uf ~., treino) |>
                    step_scale(all_numeric_predictors())
model_tree <- decision_tree(mode = "classification", engine = "rpart")
wf_tree <- workflow(preparation_tree, model_tree)
fitted_model_tree <- fit(wf_tree, treino)
#
#
predicoes <- predict(fitted_model_tree, teste)
predicoes <- mutate(predicoes, y = teste$uf)
metric <- predicoes |> metrics(y, .pred_class)
print(metric)
```

Figure 4. Preparação dos dados, criação e aplicação do modelo

4. Experimentos e Resultados

Como relatado, a base de dados escolhida tem 483.759 informações, as configurações da máquina utilizada para os experimentos são:

- Processador: Intel core i7 de 2,6 GHz
- Memória: 16GB de 2667 MHz
- Sistema Operacional: macOS Monterey

Como a tarefa considerada mais díficil é criar as regras de decisão da nossa árvore, dedicamos uma porção maior do nosso conjunto de dados para o processo de treino. Assim 20% foi usado no processo de validação (teste) do modelo contruído, e 80% foi destinado para o treino. Nossa acurácia se manteve perto de 21%, não chegando à um resultado satisfatório.

5. Considerações Finais

Neste trabalho, escolhemos um conjunto de dados governamental com característica de consumo de energia elétrica por região, nosso objetivo era classificar o estado a qual uma amostra pertence, de acordo com os seguintes atributos: região, tipo de consumo, especificação do tipo de consumo, e nível de consumo energético.

Para fazer essa classificação utilizamos o algoritmo de aprendizagem de máquina de Árvore de Decisão. Não alcançamos o resultado esperado, devido a baixa acurácia do nosso modelo. Como trabalho futuro queremos averiguar a influência do algoritmo de k-vizinhos mais próximos na acurácia do problema proposto.

References

[gov.br] gov.br. Vantagens da publicação de dados abertos. `https://kit.dados.gov.br/vantagens-dados-abertos/`.