Enap

Análise de Dados em Linguagem R

Módulo 4

Análise de Dados na Prática

Fundação Escola Nacional de Administração Pública

Presidente

Diogo Godinho Ramos Costa

Diretor de Desenvolvimento Profissional

Paulo Marques

Coordenador-Geral de Educação a Distância

Carlos Eduardo dos Santos

Equipe responsável

Ana Carla Gualberto Cardoso (Diagramação, 2020).

Ana Paula Medeiros Araújo (Direção e Produção Gráfica, 2020).

Douglas Gomes Ferreira (Conteudista, 2020).

Guilherme Teles da Mota (Implementação Rise, 2020).

Iara da Paixão Corrêa Teixeira (Designer Instrucional, 2020).

Juliana Bermudez Souto de Oliveira (Revisão Textual, 2020).

Larisse Padua da Silva (Produção Audiovisual, 2020).

Michelli Batista Lopes (Produção Audiovisual e Implementação, 2020).

Patrick Coelho (Implementação Moodle, 2020).

Sheila Rodrigues de Freitas (Coordenação Web, 2020).

Desenvolvimento do curso realizado no âmbito do acordo de Cooperação Técnica FUB / CDT / Laboratório Latitude e Enap.

Curso produzido em Brasília, 2020.

Enap Escola Nacional de Administração Pública

Enap, 2020

Enap Escola Nacional de Administração Pública

Diretoria de Educação Continuada

SAIS - Área 2-A - 70610-900 — Brasília, DF

Sumário

Unidade 1 - Aplicação da Linguagem R na Análise de Dados 5

1.1 Analisando dados abertos de viagens a serviço..... 5

Unidade 2 - Aplicação da Linguagem R na Análise de Dados 7

2.1 Prevendo a ocorrência de diabetes 7

Referências..... 10

Módulo 4 Análise de Dados na Prática

DESTAQUE

Ao final deste módulo, você deverá ser capaz de compreender as etapas do processo de análise de dados na prática. Serão abordados dois exemplos: no primeiro, será feita a análise de dados abertos de diárias e passagens; no segundo, será construído um modelo preditivo para a prevenção da ocorrência de diabetes.

Unidade 1 - Aplicação da Linguagem R na Análise de Dados

1.1 Analisando dados abertos de viagens a serviço

A proposta deste tópico é colocar em prática algumas das funções do R trabalhando com a análise de dados abertos de viagens a serviço, com o intuito de subsidiar a tomada de medidas mais eficientes na redução dos gastos com os custos dessas viagens no setor público.

Para alcançarmos o nosso objetivo, precisamos seguir algumas etapas básicas:

➢ **Definição do problema**

Para resolver um problema, primeiramente temos que entendê-lo. Assim, precisamos entender os gastos com viagens a serviço para tomar medidas mais eficientes e, com isso, reduzir os custos dessas viagens.

Vamos levantar algumas questões relevantes acerca dessa problemática:

- Qual é o valor gasto por órgão?
- Qual é o valor gasto por cidade?
- Qual é a quantidade de viagens por mês?

Será que é possível responder esses questionamentos por meio da análise de dados utilizando a linguagem R? É o que aprenderemos na próxima etapa.

➢ **Obtenção dos dados**

Para entender como obter os dados que vamos usar na análise, acesse a demonstração contida no vídeo a seguir:

Vídeo 16: Obtendo os dados

Normalmente, as bases de dados disponíveis na internet possuem uma usabilidade simples e são muito intuitivas. Dessa forma, fica fácil coletar os dados e baixá-los para treinar algumas análises. Que tal embarcar nessa aventura?

➢ **Transformação dos dados obtidos**

É importante ter em mente que, na linguagem R, diversas transformações nos dados coletados podem ser realizadas, a depender do tipo de dado e do objetivo da análise.

O vídeo a seguir demonstra uma transformação nos dados muito utilizada nas datas coletadas, uma vez que a disposição do formato das datas pode variar de um país para outro. Acompanhe:

Vídeo 17: Transformando os dados

Existem vários formatos de datas disponíveis para serem utilizados. Se desejar consultar outros formatos, acesse o endereço: https://www.statmethods.net/input/dates.html.

➢ **Exploração dos dados**

A linguagem R possui diversas funções prontas que podem nos ajudar na exploração dos dados.

Acompanhe algumas dessas funções no vídeo a seguir:

Vídeo 18: Explorando os dados

➢ **Visualização dos resultados**

A visualização dos resultados é a etapa final do nosso estudo acerca da análise dos dados de viagens a serviço. Assim, é o momento de responder às questões levantadas na primeira fase: definição do problema. Vamos relembrá-las?

- Qual é o valor gasto por órgão?
- Qual é o valor gasto por cidade?
- Qual é a quantidade de viagens por mês?

O vídeo a seguir demonstra como obter a resposta para esses questionamentos.

Vídeo 19: Visualizando os resultados parte 1

Pronto, conseguimos responder todas as perguntas levantadas inicialmente. Você aprendeu como definir o problema, como obter os dados, como tratá-los e como utilizá-los para responder as perguntas de negócio.

Até esse momento, mostramos a análise de dados obtendo como resultado tabelas e gráficos. Agora, ensinaremos outra forma de apresentar os resultados da sua análise utilizando o R Markdown.

O R Markdown, também conhecido como Markdown, é uma ferramenta usada para transformar as análises em documentos, relatórios, apresentações e até mesmo *dashboards* de alta qualidade e de maneira programática.

O vídeo a seguir demonstra como utilizar essa ferramenta. Acompanhe:

Vídeo 20: Visualizando os resultados parte 2

Dessa forma, concluímos a análise dos dados abertos de viagens a serviço e aprendemos como gerar relatórios das análises efetuadas em diferentes formatos.

Esperamos que esse conhecimento possa facilitar seu trabalho com os dados de seu órgão e permitir a elaboração de relatórios mais efetivos.

Unidade 2 - Aplicação da Linguagem R na Análise de Dados

2.1 Prevendo a ocorrência de diabetes

Neste tópico, aprenderemos as etapas de construção de um modelo de *machine learning* utilizando a linguagem R. O objetivo dessa análise é demonstrar a construção do modelo preditivo, usado para identificar padrões e mostrar o que pode acontecer de acordo com os dados analisados.

As etapas para a construção do modelo preditivo são:

- **Definição do problema**

Geralmente, iniciamos uma análise de dados diante de um problema a ser resolvido, que requer uma análise dos dados e das informações existentes acerca da temática tratada.

A definição do problema irá guiar todas as etapas para a construção do modelo preditivo, possibilitando a apresentação dos resultados de forma adequada.

Com isso, a frase que expressa a definição do nosso problema é:

Identificar pacientes com alta probabilidade de serem diagnosticados com diabetes, tendo, no mínimo, 75% de acurácia.

Uma vez definido o problema, podemos prosseguir para a próxima fase.

➢ **Obtenção dos dados**

A fim de obter os dados, precisamos decidir qual será a fonte usada para a extração. Nesse caso, os dados utilizados estão disponíveis na biblioteca do curso. O arquivo desses dados é um *dataset* público com características de pessoas que desenvolveram e de pessoas que não desenvolveram diabetes.

Primeiro, é necessário acessar esse arquivo e fazer o download. Em seguida, deve-se carregar os dados no R. Acompanhe no vídeo a seguir como realizar esse último passo.

Vídeo 21: Obtenção dos dados

➢ **Preparação dos dados**

A linguagem R define automaticamente o tipo dos dados, mas é aconselhável verificá-los a fim de se trabalhar da melhor forma possível.

O vídeo a seguir demonstra como preparar os dados em R.

Vídeo 22: Preparação dos dados

➢ **Análise exploratória**

Esta fase é muito importante, pois possibilita a extração de informações relevantes sobre um conjunto de dados.

Acompanhe no vídeo a seguir:

Vídeo 23: Análise Exploratória

Após compreender essas etapas, vamos estudar a construção do modelo preditivo.

➢ **Construção do modelo**

Nesta etapa, você aprenderá como construir um modelo preditivo a partir dos dados públicos de pessoas que desenvolveram e que não desenvolveram a doença diabetes.

Acompanhe no vídeo a seguir:

Vídeo 24: Construção do modelo

Seguindo o passo a passo demonstrado, você conseguirá construir muitos outros modelos com outras bases de dados.

- **Visualização dos resultados**

A etapa final é a visualização dos resultados.

O vídeo a seguir aborda a entrega dos resultados da análise realizada com a criação de um produto, por exemplo, um relatório. Acompanhe:

Vídeo 25: Visualização dos Resultados

Chegamos ao final deste curso e, com isso, esperamos ter atingido a missão de apresentar um panorama do processo de análise de dados por meio da linguagem R. Além disso, almejamos ter despertado a curiosidade e a motivação para avançar nos estudos em uma temática considerada tão árida para muitos cursistas.

Referências

Unidade 1 - Aplicação da Linguagem R na Análise de Dados

BRASIL. Controladoria-Geral da União. **Portal da Transparência**. Brasília: CGU, c2020. Disponível em: http://portaltransparencia.gov.br/. Acesso em: 7 ago. 2020.

DUTT, A. Splitting a data frame into training and testing sets in R. **Stories Data Speak**, Kuala Lumpur, 6 abr. 2015. Disponível em: https://duttashi.github.io/blog/splitting-a-data-frame-into-training-and-testing-sets-in-r/. Acesso em: 11 ago. 2020.

JAWAHARLAL, V. KNN Using caret R package. **R Pubs**, [*s. l.*], 29 abr. 2014. Disponível em: https://rpubs.com/njvijay/16444. Acesso em: 11 ago. 2020.

KABACOFF, R. I. Date Values. **Quick-R**, [*s. l.*], c2017.Disponível em: https://www.statmethods.net/input/dates.html. Acesso em: 11 ago. 2020.

KASSAMBARA, A. Ggplot2 barplots: Quick start guide - R software and data visualization. **Statistical tools for high-throughput data analysis**, [*s. l.*], [2020]. Disponível em: http://www.sthda.com/english/wiki/ggplot2-barplots-quick-start-guide-r-software-and-data-visualization. Acesso em: 11 ago. 2020.

KASSAMBARA, A. Ggplot2 line plot: Quick start guide - R software and data visualization. **Statistical tools for high-throughput data analysis**, [*s. l.*], [2020]. Disponível em: http://www.sthda.com/english/wiki/ggplot2-line-plot-quick-start-guide-r-software-and-data-visualization. Acesso em: 11 ago. 2020.

KASSAMBARA, A. KNN: K-Nearest Neighbors Essentials. **Statistical tools for high-throughput data analysis**, [*s. l.*], 11 mar. 2018. Disponível em: http://www.sthda.com/english/articles/35-statistical-machine-learning-essentials/142-knn-k-nearest-neighbors-essentials/. Acesso em: 11 ago. 2020.

KUHN, M. **The Caret Package**. [*S. l.: s. n.*], 27 mar. 2019. Disponível em: http://topepo.github.io/caret/index.html. Acesso em: 11 ago. 2020.

R MARKDOWN Cheat Sheet. Boston: R Studio, 2015. Disponível em: https://rstudio.com/wp-content/uploads/2015/02/rmarkdown-cheatsheet.pdf. Acesso em: 11 ago. 2020.

TOTH, M. Detailed Guide to the Bar Chart in R with ggplot. **R-bloggers**, [*s. l.*], 1 maio 2019. Disponível em:https://www.r-bloggers.com/detailed-guide-to-the-bar-chart-in-r-with-ggplot/. Acesso em: 10 ago. 2020.

XIE, Y.; ALLAIRE, J. J.; GROLEMUND, G. **R Markdown**: The Definitive Guide. Boca Ratón: CRC Press, 2020. Disponível em: https://bookdown.org/yihui/rmarkdown/installation.html. Acesso em: 11 ago. 2020.

TUNEGRID and TuneLength in Caret. **R Pubs**, [*s. l.*], [2018]. Disponível em: https://rpubs.com/Mentors_Ubiqum/tunegrid_tunelength. Acesso em: 11 ago. 2020.

Unidade 2 - Aplicação da Linguagem R na Análise de Dados

BRASIL. Controladoria-Geral da União. **Portal da Transparência**. Brasília: CGU, c2020. Disponível em: http://portaltransparencia.gov.br/. Acesso em: 7 ago. 2020.

DUTT, A. Splitting a data frame into training and testing sets in R. **Stories Data Speak**, Kuala Lumpur, 6 abr. 2015. Disponível em: https://duttashi.github.io/blog/splitting-a-data-frame-into-training-and-testing-sets-in-r/. Acesso em: 11 ago. 2020.

JAWAHARLAL, V. KNN Using caret R package. **R Pubs**, [*s. l.*], 29 abr. 2014. Disponível em: https://rpubs.com/njvijay/16444. Acesso em: 11 ago. 2020.

KABACOFF, R. I. Date Values. **Quick-R**, [*s. l.*], c2017.Disponível em: https://www.statmethods.net/input/dates.html. Acesso em: 11 ago. 2020.

KASSAMBARA, A. Ggplot2 barplots : Quick start guide - R software and data visualization. **Statistical tools for high-throughput data analysis**, [*s. l.*], [2020]. Disponível em: http://www.sthda.com/english/wiki/ggplot2-barplots-quick-start-guide-r-software-and-data-visualization. Acesso em: 11 ago. 2020.

KASSAMBARA, A. Ggplot2 line plot: Quick start guide - R software and data visualization. **Statistical tools for high-throughput data analysis**, [*s. l.*], [2020]. Disponível em: http://www.sthda.com/english/wiki/ggplot2-line-plot-quick-start-guide-r-software-and-data-visualization. Acesso em: 11 ago. 2020.

KASSAMBARA, A. KNN: K-Nearest Neighbors Essentials. **Statistical tools for high-throughput data analysis**, [*s. l.*], 11 mar. 2018. Disponível em: http://www.sthda.com/english/articles/35-statistical-machine-learning-essentials/142-knn-k-nearest-neighbors-essentials/. Acesso em: 11 ago. 2020.

KUHN, M. **The Caret Package**. [*S. l.: s. n.*], 27 mar. 2019. Disponível em: http://topepo.github.io/caret/index.html. Acesso em: 11 ago. 2020.

R MARKDOWN Cheat Sheet. Boston: R Studio, 2015. Disponível em: https://rstudio.com/wp-content/uploads/2015/02/rmarkdown-cheatsheet.pdf. Acesso em: 11 ago. 2020.

TOTH, M. Detailed Guide to the Bar Chart in R with ggplot. **R-bloggers**, [*s. l.*], 1 maio 2019. Disponível em:https://www.r-bloggers.com/detailed-guide-to-the-bar-chart-in-r-with-ggplot/. Acesso em: 10 ago. 2020.

TUNEGRID and TuneLength in Caret. **R Pubs**, [*s. l.*], [2018]. Disponível em: https://rpubs.com/Mentors_Ubiqum/tunegrid_tunelength. Acesso em: 11 ago. 2020.

XIE, Y.; ALLAIRE, J. J.; GROLEMUND, G. **R Markdown**: The Definitive Guide. Boca Ratón: CRC Press, 2020. Disponível em: https://bookdown.org/yihui/rmarkdown/installation.html. Acesso em: 11 ago. 2020.