GOVERNO DO ESTADO DO ESPÍRITO SANTO
SECRETARIA DE ECONOMIA E PLANEJAMENTO – SEP
INSTITUTO JONES DOS SANTOS NEVES – IJSN

MANUAL

Preditor do Abandono Escolar

Manual do pacote em R
PreditorIJSN

Instituto Jones dos Santos Neves

GOVERNO DO ESTADO DO ESPÍRITO SANTO

Julho de 2023

GOVERNO DO ESTADO DO ESPÍRITO SANTO
José Renato Casagrande

VICE-GOVERNADORIA
Ricardo Ferraço

SECRETARIA DE ECONOMIA E PLANEJAMENTO – SEP
Álvaro Rogério Duboc Fajardo

INSTITUTO JONES DOS SANTOS NEVES – IJSN

Diretor Presidente
Pablo Silva Lira

Diretoria de Estudos e Pesquisas
Pablo Medeiros Jabor

Diretoria de Integração e Projetos Especiais
Antonio Ricardo Freislebem da Rocha

Diretoria de Gestão Administrativa
Katia Cesconeto de Paula

Coordenação Geral
Kiara de Deus Demura

Elaboração
Guilherme Armando de Almeida Pereira (Bolsista Fapes)
Kiara de Deus Demura

Instituto Jones dos Santos Neves

Sumário

1. Introdução 4

2. Instalação 4

3. Como organizar o seu banco de dados 6

4. Exemplo 7

5. Como citar 8

Referências 8

1. Introdução

O objetivo deste documento é apresentar as principais funcionalidades do preditor do abandono das escolas estaduais do Espírito Santo (**PreditorIJSN**).

A descrição dos modelos disponíveis no pacote do Preditor está disponível no Relatório – Preditor do Abandono Escolar (PEREIRA; DEMURA, 2023). Nossa ferramenta utiliza como suporte os pacotes *caret* (2022), *tidyverse* (2019), *ROCR* (2005) e *ggpubr* (2020).

Recomenda-se, como introdução aos modelos utilizados, as referências James et al. (2013) e Hastie, Tibshirani e Friedman (2009).

2. Instalação

Passo 1: Instalando os pacotes auxiliares

Para que o **PreditorIJSN** funcione corretamente é necessário que o usuário tenha instalado os seguintes pacotes auxiliares em sua máquina: $(caret, tidyverse, ROCR\ e\ ggpubr)$. Esta ação precisa ser realizada apenas uma única vez.

Para executar a instalação, execute as seguintes linhas de comando:

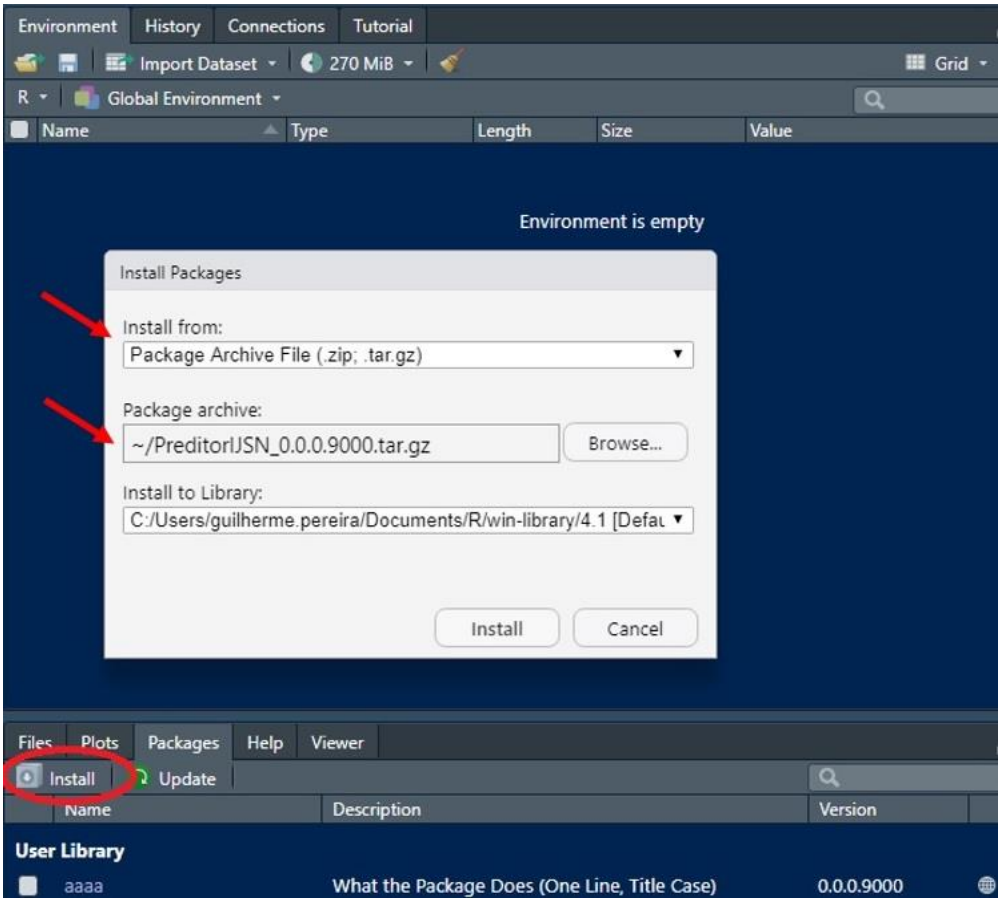
```
#
install.packages("caret")
install.packages("tidyverse")
install.packages("ROCR")
install.packages("ggpubr")
#
```

Passo 2: Instalando o PreditorIJSN

Há duas formas de se instalar o preditor. A primeira maneira pode ser realizada utilizando o *mouse* do computador. Para isso vá em *Install*. Em *Install from*, selecione "Package Archive File (.zip; .tar.gz)". Por fim, por meio do *Browse* você deve procurar o arquivo do preditor previamente baixado em seu computador.

A Figura 1 abaixo ilustra esse procedimento.

Figura 1 – Instalando o pacote PreditorIJSN



Elaboração: Estudos Educacionais (IJSN).

A segunda opção é por meio do comando:

```
#
install.packages("~/PreditorIJSN_0.0.1.0000.tar.gz", repos = NULL, typ
e = "source")
#
```

Vale a pena observar que o diretório declarado deve conter o arquivo original do $PreditorIJSN$.

Passo 3: Carregando o pacote ***PreditorIJSN***

Uma vez instalados os pacotes, devemos carregá-los por meio do seguinte comando:

```
#
library(PreditorIJSN)
#
```

3. Como organizar o seu banco de dados

O pacote disponibiliza quatro bancos de dados fictícios, sendo dois para o ensino médio e dois para o ensino fundamental. O objetivo é que o usurário possa visualizar como o seu banco deve ser elaborado. Os bancos presentes no pacote podem ser acessados ao se digitar:

```
#
data(dados2021_EF)
data(dados2022_EF)
data(dados2021_EM)
data(dados2022_EM)
#
```

O banco de dados que você irá construir ou preparar para a previsão podem conter diversas variáveis, não sendo restrito às variáveis presentes nos bancos ilustrativos. Contudo, algumas especificações devem ser respeitadas:

(i) As linhas representam os estudantes;

(ii) As colunas indicam as variáveis;

(iii) As variáveis devem ser, obrigatoriamente, do tipo *factor* para as variáveis qualitativas e do tipo *numeric* para as variáveis quantitativas;

(iv) A variável que indica o abandono (quando pertinente) deve se chamar *abandono* e possuir o seguinte código: *abandono=1* corresponde ao abandono e *abandono=0* correspondente ao não abandono;

(v) É mandatório que a variável que identifica o aluno seja nomeada como *CD_INEP_ALUNO*.

(vi) É obrigatório a presença de uma variável denominada *ID_ETAPA_MATRICULA* que indica a série em análise. Esta é uma variável do tipo *factor* com códigos variando de 1 (primeiro ano do ensino fundamental) até 9 (nono ano do ensino fundamental). Os níveis dessa variável podem ser alterados de acordo com o interesse do usuário. Caso o banco de dados seja relativo ao ensino médio, esta variável possuirá 3 níveis.

(vii) É obrigatório que o banco de dados contenha *CD_INEP_ESC* indicando o código INEP da escola.

(viii) É obrigatório que o banco contenha a variável *NOME_ESCOLA* designando o nome da escola.

Em suma, independentemente do número de variáveis que o seu banco contém, é mandatório que as seguintes variáveis estejam presentes e nomeadas exatamente como apresentadas: *abandono* (para o banco de dados utilizado para a estimação), *CD_INEP_ALUNO*, *ID_ETAPA_MATRICULA*, *CD_INEP_ESC* e *NOME_ESCOLA*. A Figura 2

abaixo ilustra como o banco deve estar organizado. É recomendado também que não haja dados faltantes (*NA*) no seu banco.

Figura 2 – Exemplo de banco de dados

	abandono	CD_INEP_ALUNO	ID_ETAPA_MATRICULA	CD_INEP_ESC	NOME_ESCOLA	IDADE	TP_SEXO	NOTA_ESCOLA_TRI1PT
1	1	1111	6	1750	Sao Jose	19	F	14.21503
2	1	1113	1	1310	Maestro Guerra Peixe	19	M	19.76923
3	1	1115	4	1860	Canarinhos	19	M	12.77670
4	1	1116	8	1310	Maestro Guerra Peixe	19	M	15.41554
5	1	1117	5	1860	Canarinhos	23	F	22.60502
6	1	1119	5	1640	Ipiranga	20	F	18.32938
7	1	1121	4	1860	Canarinhos	19	F	18.91000

Elaboração: Estudos Educacionais (IJSN).

4. Exemplo

Pronto! Uma vez instalados e carregados os pacotes, e em posse dos bancos de dados, podemos utilizar o *PreditorIJSN*. O pacote desenvolvido possui duas funções principais: uma relativa à previsão do ensino fundamental (*PreditorEnsinoFundamental*) e a outra à previsão do ensino médio (*PreditorEnsinoMedio*).

Nesse exemplo vamos estimar o modelo para o 1º ano do ensino médio. Para estimar qualquer modelo você deve ter em suas mãos:

(i) Banco de dados relativo ao ano que o modelo será estimado. Geralmente utilizamos o ano anterior ao ano que queremos fazer as previsões. Nesse exemplo vamos utilizar `dados_treinamento = dados2021_EM`.

(ii) Banco de dados relativo ao ano que iremos fazer as previsões. É importante destacar que este banco não contém a variável *abandono*, uma vez que esta será prevista. Nesse exemplo vamos utilizar `dados_treinamento = dados2022_EM`.

Estimando o modelo e fazendo as previsões:

```
model.fit <- PreditorEnsinoMedio(dados_treinamento = dados2021_EM, dad
os_previsao = dados2022_EM, ETAPA.MATRICULA = 16, TesteValidacao = TRU
E, p = 0.8, tipo="lasso", dir="D:/Local_dos_relatorios/")
```

Vale ressaltar que no diretório especificado pelo usuário `dir="D:/Local_dos_relatorios/"` o pacote irá criar dois arquivos. O primeiro arquivo possui a extensão *.csv* e contém as seguintes informações:

(i) *Status* - previsão do abandono.
(ii) *Prob_Y1* - probabilidade de abandono.
(iii) *CD_INEP_ALUNO* - cód. INEP do aluno.
(iv) *ID_ETAPA_MATRICULA* - indica a série.
(v) *CD_INEP_ESCOLA* - cód. INEP da escola.
(vi) *NOME_ESCOLA* - nome da escola.

A Figura 3 abaixo apresenta a correspondente saída *.csv*.

Figura 3 – Exemplo da saída *.csv*

Status	Prob_Y1	CD_INEP_ALUNO	ID_ETAPA_MATRICULA	CD_INEP_ESC	NOME_ESCOLA
1	0.57107351922838	1347	17	1860	Canarinhos
1	0.568771102872594	2160	17	1750	Sao Jose
0	0.44601371809005	1206	17	1310	Maestro Guerra Peixe
0	0.44295923227065	1261	17	1420	Dom Pedro II
1	0.565696600742819	2130	18	1640	Ipiranga

Elaboração: Estudos Educacionais (IJSN).

O segundo arquivo possui o formato *.pdf* e contém informações sobre a estimação assim como métricas de desempenho para o conjunto de treinamento.

5. Como citar

PEREIRA, Guilherme A. de A.; DEMURA, Kiara de Deus; NUNES, Iago C. **PreditorIJSN**. R package version 0.0.1.0000. 2023. Disponível em: <https://ijsn.es.gov.br/publicacoes/relatorios/preditor-do-abandono-escolar>." Acesso em: (*DATA_ACESSO*).

Referências

HASTIE, Trevor; TIBSHIRANI, Robert; FRIEDMAN Jerome. **The elements of statistical learning: data mining, inference and prediction.** New York: Springer series in statistics, 2009.

JAMES, Gareth et al. **An introduction to statistical learning with applications in R**. New York: Springer, 2013.

KASSAMBARA, A. **Ggpubr**: 'ggplot2' Based Publication Ready Plots. R package version 0.4.0. 2020. Disponível em: <https://CRAN.R-project.org/package=ggpubr>.

KUHN, M. **Caret**: Classification and Regression Training. R package version 6.0-82. 2022. Disponível em <https://CRAN.R-project.org/package=caret>.

SING, T. et al. ROCR: visualizing classifier performance in R. **Bioinformatics**, v. 21, n. 20, p. 3940-3941, 2005. Disponível em: <http://rocr.bioinf.mpi-sb.mpg.de>.

WICKHAM et al., (2019). Welcome to the tidyverse. **Journal of Open Source Software**, v. 4, n. 43, 2019. Disponível em: <https://doi.org/10.21105/joss.01686>.