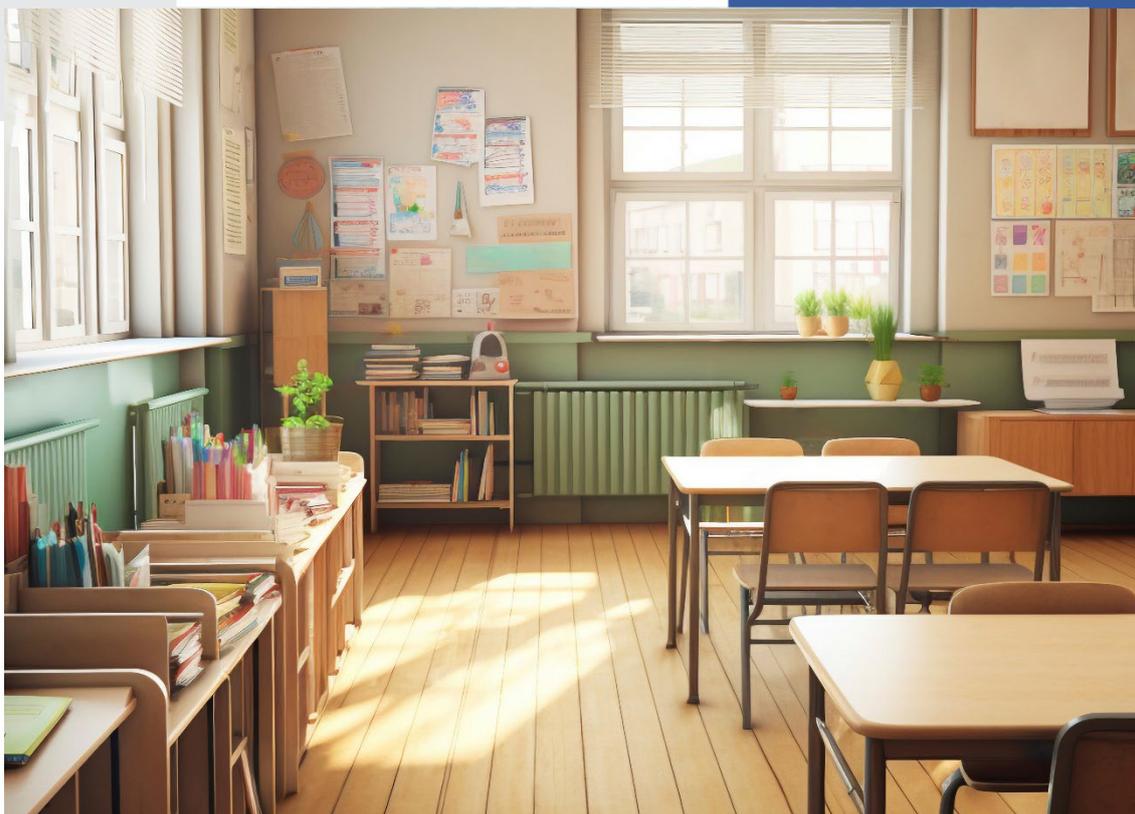


GOVERNO DO ESTADO DO ESPÍRITO SANTO  
SECRETARIA DE ECONOMIA E PLANEJAMENTO – SEP  
INSTITUTO JONES DOS SANTOS NEVES – IJSN

## MANUAL



# Preditor do Abandono Escolar

*Manual do pacote em R  
PreditorIJSN*



Instituto Jones  
dos Santos Neves



GOVERNO DO ESTADO  
DO ESPÍRITO SANTO

Dezembro de 2024

**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 – Pesquisador Bolsista (Fapes)

Kiara de Deus Demura



## Sumário

1. Introdução .....	4
2. Instalação.....	4
3. Como organizar os seus bancos de dados.....	6
4. Exemplo .....	7
Referências .....	9

## 1. Introdução

O objetivo deste documento é apresentar as principais funcionalidades do preditor do abandono das escolas estaduais do Espírito Santo (**PreditorIJSN**). Esse pacote, desenvolvido em R, é um produto da parceria entre o Instituto Jones dos Santos Neves (IJSN), Secretaria de Estado da Educação (SEDU) e Fundação de Amparo à Pesquisa e Inovação do Espírito Santo (FAPES) – Estudos Educacionais, a fim de auxiliar a gestão da SEDU na prevenção ao abandono escolar.

No total cinco modelos distintos estão disponíveis: i) regressão logística com lasso; (ii) regressão logística com *ridge*; (iii) regressão logística com *elastic-net*; (iv) naive Bayes Gaussiano e, (v) redes neurais artificiais (*feedforward neural network*). Maiores detalhes sobre a formulação matemática dessas ferramentas podem ser encontrados em Hastie, Tibshirani e Friedman (2009), Venables e Ripley (2002) e James et al. (2013).

Adicionalmente, dois métodos de pré-processamento para balanceamento da amostra estão disponíveis: (i) *downsample* e (ii) *synthetic minority oversampling technique* (SMOTE). Neste caso, recomenda-se as referências Chawla et al. (2002) e Fernández et al. (2018).

Esta ferramenta utiliza como suporte os pacotes *caret* (KUNH, 2008), *tidyverse* (WICKHAM et al., 2019), *ROCR* (SING et al. 2005), *performanceEstimation* (TORGO, 2014) e *ggpubr* (KASSAMBARA, 2023).

A seguir, são apresentados os procedimentos para instalação, considerando a utilização do ambiente de desenvolvimento integrado (*integrated development environment* – IDE) *RStudio*, bem como as instruções para a organização dos bancos de dados. Por fim, um exemplo é realizado para previsão de uma série hipotética do ensino médio.

## 2. Instalação

### *Passo 1: Instalando os pacotes auxiliares*

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

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

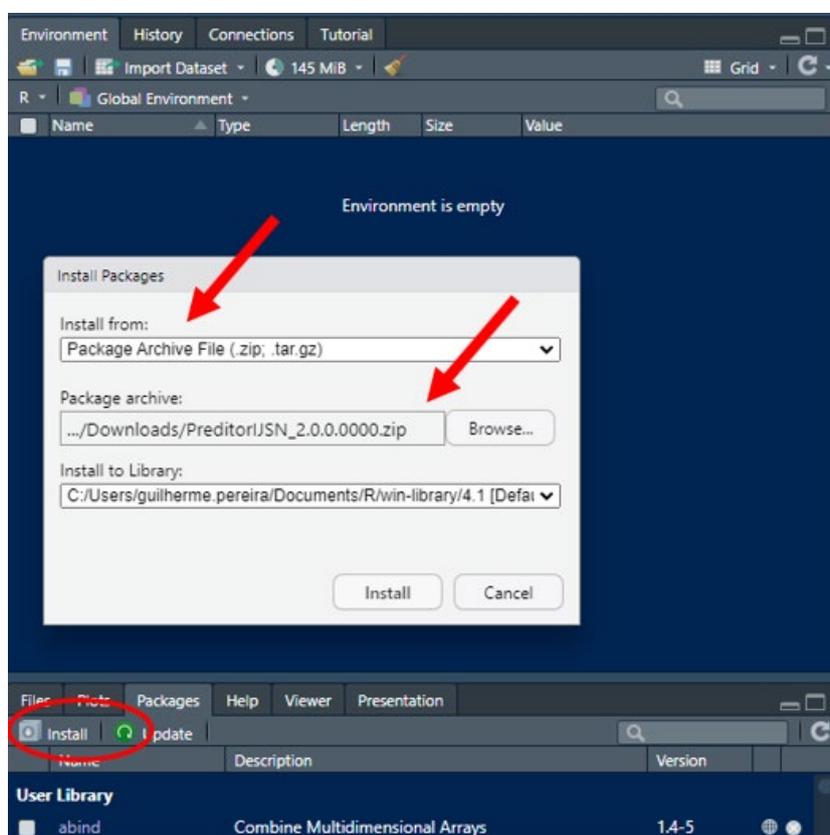
```
#  
install.packages("caret")  
install.packages("tidyverse")  
install.packages("ROCR")
```

```
install.packages("performanceEstimation")  
install.packages("ggpubr")  
#
```

### Passo 2: Instalando o *PreditorIJSN*

Há duas formas de instalar o preditor. A primeira maneira é com o auxílio do *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 em seu computador. A Figura 1 ilustra esse procedimento.

Figura 1 – Instalando o pacote *PreditorIJSN*



Elaboração: Estudos Educacionais/IJSN.

A segunda opção utiliza o comando:

```
#  
install.packages("~/Local_do_pacote/PreditorIJSN_3.0.0.0000.zip", repos = NULL, type = "source")  
#
```



Vale a pena observar que o diretório declarado na função `install.packages("~/Local_do_pacote/...")` 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 os seus bancos de dados

O pacote disponibiliza quatro bancos de dados fictícios, sendo dois para o ensino médio e dois para o ensino fundamental. Assim, o usuário pode visualizar como os seus bancos devem ser elaborados. Os bancos presentes no pacote podem ser acessados ao digitar:

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

Os bancos de dados para a previsão podem conter diversas variáveis, não sendo restritos 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 ser denominada como *abandono* e possuir o seguinte código: (i) *abandono=1*, corresponde ao abandono; (ii) *abandono=0*, corresponde ao não abandono;
- (v) É mandatório que a variável que identifica o aluno seja nomeada como *CD\_INEP\_ALUNO*;
- (vi) É obrigatória a presença de uma variável denominada *ID\_ETAPA\_MATRICULA*, para indicar a série em análise. Esta é uma variável do tipo *factor* com códigos variando de 1 (1º ano do ensino fundamental) até 9 (9º 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 a variável *CD\_INEP\_ESC* indicando o código INEP da escola ou equivalente;
- (viii) É obrigatório que o banco contenha a variável *NOME\_ESCOLA* designando o nome da escola.

Em suma, independentemente do número de variáveis que os bancos de dados possuam, é mandatório que as seguintes variáveis estejam presentes e nomeadas exatamente como apresentadas: *abandono* (apenas para o banco de dados utilizado na estimação), *CD\_INEP\_ALUNO*, *ID\_ETAPA\_MATRICULA*, *CD\_INEP\_ESC* e *NOME\_ESCOLA*.

A Figura 2 ilustra como os bancos devem estar organizados. É recomendado também que não haja dados faltantes (*NA*) em seus bancos, pois o pacote não realiza qualquer procedimento para preenchimento dos mesmos.

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 preditor. O pacote desenvolvido possui duas funções principais: uma relativa à previsão para o ensino fundamental (*PreditorEnsinoFundamental*) e a outra à previsão para o ensino médio (*PreditorEnsinoMedio*).

Nesse exemplo vamos estimar o modelo para a 1ª série 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. Neste exemplo vamos utilizar `dados_previsao = dados2022_EM`.

Estimando o modelo e fazendo as previsões:

```
#  
model.fit <- PreditorEnsinoMedio(dados_treinamento = dados2021_EM, dados_previsao = dados2022_EM, ETAPA.MATRICULA = 16, TesteValidacao = TRUE, p = 0.8, tipo="lasso", balanceamento = "smote", dir="D:/Local_dos_relatorios/")  
#
```

Vale ressaltar que no diretório especificado pelo usuário em `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 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.

## Referências

CHAWLA, Nitesh V. et al. SMOTE: Synthetic Minority Over-sampling Technique. **Journal of Artificial Intelligence Research**, v. 16, p. 321-357, 2002.

FERNÁNDEZ, Alberto et al. SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary. **Journal of Artificial Intelligence Research**, v. 61, p. 863-905, 2018.

FRIEDMAN, Jerome; HASTIE, Trevor; TIBSHIRANI, Rob. Regularization Paths for Generalized Linear Models via Coordinate Descent. **Journal of Statistical Software**, v. 33, n. 1, p. 1, 2010. DOI 10.18637/jss.v033.i01. Disponível em: <https://doi.org/10.18637/jss.v033.i01>.

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**. New York: Springer, 2013.

KASSAMBARA, A. ggpubr: 'ggplot2' Based Publication Ready Plots. R package version 0.6.0. 2023. Disponível em: <https://CRAN.R-project.org/package=ggpubr>.

KUHN, Max. Building predictive models in R using the caret package. **Journal of Statistical Software**, v. 28, p. 1-26, 2008. DOI 10.18637/jss.v028.i05. Disponível em: <https://doi.org/10.18637/jss.v028.i05>.

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>.

TORGO, Luis. An infra-structure for performance estimation and experimental comparison of predictive models in R. **CoRR**, 2014. Disponível em: <https://arxiv.org/abs/1412.0436>.

VENABLES, W. N.; RIPLEY, B. D. **Modern Applied Statistics with S**. 4. ed. New York: Springer, 2002.

WEIHS, Claus et al. klaR Analyzing German Business Cycles. *In*: Baier, D.; Decker, R.; SCHMIDT-THIEME, L. **Data Analysis and Decision Support**. Berlin: Springer-Verlag, 2005. p. 335-343.

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