



UNIVERSIDADE FEDERAL DO RIO DE JANEIRO

ESCOLA POLITÉCNICA

DEPARTAMENTO DE ELETRÔNICA E DE

COMPUTAÇÃO

# Introdução ao Aprendizado de Máquina

## Trabalho 1

Classificador para aprovação de crédito

*Juan Coutinho Lima*

21 de maio de 2021

# Sumário

<b>Lista de Figuras</b>	<b>1</b>
<b>Lista de Tabelas</b>	<b>1</b>
<b>1 Introdução</b>	<b>2</b>
<b>2 Estudo da base de dados</b>	<b>2</b>
2.1 Variáveis categóricas . . . . .	2
2.1.1 Descarte das variáveis categóricas . . . . .	3
2.2 <i>One-hot encoding</i> . . . . .	3
2.3 Binarização . . . . .	4
2.4 Visualização dos dados . . . . .	4

## Lista de Figuras

1	Exemplo de One-hot encoding . . . . .	3
2	One-hot encoding regiões . . . . .	3
3	Exemplo de binarização . . . . .	4

## Lista de Tabelas

1	Variáveis categóricas . . . . .	2
---	---------------------------------	---

# 1 Introdução

O objetivo deste trabalho é construir um classificador para apoio à decisão de aprovação de crédito, com base em dados disponibilizados para a realização deste exercício. A ideia é que, através da consideração deste histórico disposto, seja possível avaliar se um cliente que executa o pedido de crédito apresentará ou não inadimplência. Para isso, precisamos realizar um estudo sobre as características da nossa base de dados, identificar os possíveis parâmetros relevantes para o classificador e assim definir o melhor qual o modelo que melhor atende os nossos requisitos.

## 2 Estudo da base de dados

A nossa base de dados consiste em um arquivo CSV com 41 atributos, sendo o parâmetro “inadimplente” a nossa variável alvo.

### 2.1 Variáveis categóricas

Para que seja possível avaliar a adequação dos nossos atributos, devemos primeiro atestar que todos eles se tratam de variáveis numéricas. O primeiro passo a ser feito é identificar quais são as variáveis categóricas. Isso pode ser realizado ao listar todas as colunas do tipo “object”.

Variável	Tipo	Número categorias
forma_envio_solicitacao	não ordinal	3
sexo	não ordinal	4
estado_onde_nasceu	não ordinal	28
estado_onde_reside	não ordinal	27
possui_telefone_residencial	binário	2
codigo_area_telefone_residencial	não ordinal	75
possui_telefone_celular	binário	2
vinculo_formal_com_empresa	binário	2
estado_onde_trabalha	não ordinal	28
possui_telefone_trabalho	binário	2
codigo_area_telefone_trabalho	não ordinal	77

Tabela 1: Variáveis categóricas

### 2.1.1 Descarte das variáveis categóricas

O objetivo desta seção é identificar quais foram as variáveis descartadas e o motivo desta decisão.

Os atributos “forma\_envio\_solicitacao”, “possui\_telefone\_residencial” e “possui\_telefone\_celular” foram considerados irrelevantes, pois a natureza destes atributos não nos permite fazer a correlação com a inadimplência. Foi considerado que os atributos “codigo\_area\_telefone\_residencial”, “estado\_onde\_trabalha” e “codigo\_area\_telefone\_trabalho” apresentam informação sobre localidade de forma redundante e portanto serão descartados.

## 2.2 One-hot encoding

Agora que as variáveis relevantes foram separadas, nos resta atribuir um tipo numérico para os seus campos. Essa atribuição de valor será feita através da técnica de *one-hot encoding*. Tomemos o atributo “sexo” como exemplo. Para esta coluna temos 4 categorias possíveis: “M”, “F”, “N”, e vazio. A primeira ação será transformar todos os campos vazios em “N”, com o uso da técnica de *one-hot encoding* teremos o seguinte resultado:

	sexo_F	sexo_M	sexo_N
18172	1	0	0
16075	1	0	0
15388	1	0	0
13842	0	1	0
8610	1	0	0

Figura 1: Exemplo de One-hot encoding

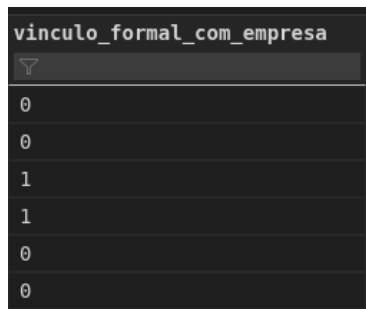
A mesma técnica foi aplicada ao atributo “estado\_onde\_reside”, com a diferença de que antes foi necessário agrupar os estados em regiões, uma pequena amostra desta operação pode ser vista na figura abaixo.

regiao_centro-oeste	regiao_nordeste	regiao_norte	regiao_sudeste	regiao_sul
0	1	0	0	0
1	0	0	0	0
1	0	0	0	0
0	0	0	1	0
0	0	0	1	0
0	1	0	0	0
0	1	0	0	0
0	0	0	1	0

Figura 2: One-hot encoding regiões

## 2.3 Binarização

A ultima parte do tratamento das variáveis categóricas consiste em transformar as variáveis com apenas duas categorias em campos com 1 e 0. Um exemplo da necessidade dessa aplicação é o atributo “vinculo\_formal\_com\_empresa” que possui os valores “Y” e “N”. Com a aplicação da técnica obtivemos o seguinte resultado:



vinculo_formal_com_empresa
Y
0
0
1
1
0
0

Figura 3: Exemplo de binarização

## 2.4 Visualização dos dados