Carregando…

classificacao\_clientes.ipynb\_

Arquivo

Editar

Ver

Inserir

Ambiente de execução

Ferramentas

Ajuda

Salvo pela última vez às 11 de agosto

ComentárioCompartilhar

[](https://accounts.google.com/SignOutOptions?hl=pt-BR&continue=https://colab.research.google.com/drive/1WZ-GEGRX9RmoiL0BXo23K1HJ_y9U4MRY)

CódigoTextoConectarEditar

cod\_cliente --> Código do cliente no conglomerado

sexo --> Sexo

estado\_civil --> Solteiro/casado

dependentes --> Quantidade de filhos

educacao --> Qual a escolaridade

empregado --> Se está trabalhando

renda --> Valor da renda mensal

renda\_conjuge --> Valor da renda mensal do companheiro

emprestimo --> Valor do empréstimo solicitado (em milhar)

prestacao\_mensal --> Valor da prestação mensal a ser paga

historico\_credito --> Se o cliente já atrasou ou esteve em situação de descumprimento das obrigações

imovel --> Se o cliente possui imóvel

aprovacao\_emprestimo --> Se houve aprovação (sim ou não) do empréstimo solicitado.

[ ]

## Pacotes

import pandas as pd     #Biblioteca de Funções de Manipulação de Dados

import numpy as np

!pip install dataprep

import seaborn as sns   #Biblioteca de Funções de Visualização Gráfica

import warnings

warnings.filterwarnings("ignore")





Collecting dataprep

Downloading dataprep-0.3.0-py3-none-any.whl (1.8 MB)

|████████████████████████████████| 1.8 MB 5.2 MB/s

Requirement already satisfied: bottleneck<2.0,>=1.3 in /usr/local/lib/python3.7/dist-packages (from dataprep) (1.3.2)

Collecting regex<2021.0.0,>=2020.10.15

Downloading regex-2020.11.13-cp37-cp37m-manylinux2014\_x86\_64.whl (719 kB)

|████████████████████████████████| 719 kB 70.4 MB/s

Collecting pydantic<2.0,>=1.6

Downloading pydantic-1.8.2-cp37-cp37m-manylinux2014\_x86\_64.whl (10.1 MB)

|████████████████████████████████| 10.1 MB 33.5 MB/s

Collecting dask[array,dataframe,delayed]<3.0,>=2.25

Downloading dask-2.30.0-py3-none-any.whl (848 kB)

|████████████████████████████████| 848 kB 48.2 MB/s

Collecting wordcloud<2.0,>=1.8

Downloading wordcloud-1.8.1-cp37-cp37m-manylinux1\_x86\_64.whl (366 kB)

|████████████████████████████████| 366 kB 34.0 MB/s

Requirement already satisfied: bokeh<3,>=2 in /usr/local/lib/python3.7/dist-packages (from dataprep) (2.3.3)

Requirement already satisfied: pandas<2.0,>=1.1 in /usr/local/lib/python3.7/dist-packages (from dataprep) (1.1.5)

Requirement already satisfied: ipywidgets<8.0,>=7.5 in /usr/local/lib/python3.7/dist-packages (from dataprep) (7.6.3)

Collecting jsonpath-ng<2.0,>=1.5

Downloading jsonpath\_ng-1.5.3-py3-none-any.whl (29 kB)

Requirement already satisfied: numpy<2,>=1 in /usr/local/lib/python3.7/dist-packages (from dataprep) (1.19.5)

Collecting tqdm<5.0,>=4.48

Downloading tqdm-4.62.0-py2.py3-none-any.whl (76 kB)

|████████████████████████████████| 76 kB 5.2 MB/s

Requirement already satisfied: scipy<2,>=1 in /usr/local/lib/python3.7/dist-packages (from dataprep) (1.4.1)

Collecting usaddress<0.6.0,>=0.5.10

Downloading usaddress-0.5.10-py2.py3-none-any.whl (63 kB)

|████████████████████████████████| 63 kB 2.8 MB/s

Collecting nltk<4.0,>=3.5

Downloading nltk-3.6.2-py3-none-any.whl (1.5 MB)

|████████████████████████████████| 1.5 MB 35.8 MB/s

Collecting aiohttp<4.0,>=3.6

Downloading aiohttp-3.7.4.post0-cp37-cp37m-manylinux2014\_x86\_64.whl (1.3 MB)

|████████████████████████████████| 1.3 MB 42.6 MB/s

Collecting metaphone<0.7,>=0.6

Downloading Metaphone-0.6.tar.gz (14 kB)

Requirement already satisfied: jinja2<3.0,>=2.11 in /usr/local/lib/python3.7/dist-packages (from dataprep) (2.11.3)

Collecting python-Levenshtein<0.13.0,>=0.12.2

Downloading python-Levenshtein-0.12.2.tar.gz (50 kB)

|████████████████████████████████| 50 kB 6.9 MB/s

Requirement already satisfied: chardet<5.0,>=2.0 in /usr/local/lib/python3.7/dist-packages (from aiohttp<4.0,>=3.6->dataprep) (3.0.4)

Requirement already satisfied: attrs>=17.3.0 in /usr/local/lib/python3.7/dist-packages (from aiohttp<4.0,>=3.6->dataprep) (21.2.0)

Requirement already satisfied: typing-extensions>=3.6.5 in /usr/local/lib/python3.7/dist-packages (from aiohttp<4.0,>=3.6->dataprep) (3.7.4.3)

Collecting multidict<7.0,>=4.5

Downloading multidict-5.1.0-cp37-cp37m-manylinux2014\_x86\_64.whl (142 kB)

|████████████████████████████████| 142 kB 65.1 MB/s

Collecting yarl<2.0,>=1.0

Downloading yarl-1.6.3-cp37-cp37m-manylinux2014\_x86\_64.whl (294 kB)

|████████████████████████████████| 294 kB 48.5 MB/s

Collecting async-timeout<4.0,>=3.0

Downloading async\_timeout-3.0.1-py3-none-any.whl (8.2 kB)

Requirement already satisfied: tornado>=5.1 in /usr/local/lib/python3.7/dist-packages (from bokeh<3,>=2->dataprep) (5.1.1)

Requirement already satisfied: PyYAML>=3.10 in /usr/local/lib/python3.7/dist-packages (from bokeh<3,>=2->dataprep) (3.13)

Requirement already satisfied: packaging>=16.8 in /usr/local/lib/python3.7/dist-packages (from bokeh<3,>=2->dataprep) (21.0)

Requirement already satisfied: python-dateutil>=2.1 in /usr/local/lib/python3.7/dist-packages (from bokeh<3,>=2->dataprep) (2.8.1)

Requirement already satisfied: pillow>=7.1.0 in /usr/local/lib/python3.7/dist-packages (from bokeh<3,>=2->dataprep) (7.1.2)

Requirement already satisfied: toolz>=0.8.2 in /usr/local/lib/python3.7/dist-packages (from dask[array,dataframe,delayed]<3.0,>=2.25->dataprep) (0.11.1)

Collecting fsspec>=0.6.0

Downloading fsspec-2021.7.0-py3-none-any.whl (118 kB)

|████████████████████████████████| 118 kB 38.5 MB/s

Collecting partd>=0.3.10

Downloading partd-1.2.0-py3-none-any.whl (19 kB)

Requirement already satisfied: cloudpickle>=0.2.2 in /usr/local/lib/python3.7/dist-packages (from dask[array,dataframe,delayed]<3.0,>=2.25->dataprep) (1.3.0)

Requirement already satisfied: ipython>=4.0.0 in /usr/local/lib/python3.7/dist-packages (from ipywidgets<8.0,>=7.5->dataprep) (5.5.0)

Requirement already satisfied: traitlets>=4.3.1 in /usr/local/lib/python3.7/dist-packages (from ipywidgets<8.0,>=7.5->dataprep) (5.0.5)

Requirement already satisfied: widgetsnbextension~=3.5.0 in /usr/local/lib/python3.7/dist-packages (from ipywidgets<8.0,>=7.5->dataprep) (3.5.1)

Requirement already satisfied: nbformat>=4.2.0 in /usr/local/lib/python3.7/dist-packages (from ipywidgets<8.0,>=7.5->dataprep) (5.1.3)

Requirement already satisfied: jupyterlab-widgets>=1.0.0 in /usr/local/lib/python3.7/dist-packages (from ipywidgets<8.0,>=7.5->dataprep) (1.0.0)

Requirement already satisfied: ipykernel>=4.5.1 in /usr/local/lib/python3.7/dist-packages (from ipywidgets<8.0,>=7.5->dataprep) (4.10.1)

Requirement already satisfied: jupyter-client in /usr/local/lib/python3.7/dist-packages (from ipykernel>=4.5.1->ipywidgets<8.0,>=7.5->dataprep) (5.3.5)

Requirement already satisfied: pexpect in /usr/local/lib/python3.7/dist-packages (from ipython>=4.0.0->ipywidgets<8.0,>=7.5->dataprep) (4.8.0)

Requirement already satisfied: setuptools>=18.5 in /usr/local/lib/python3.7/dist-packages (from ipython>=4.0.0->ipywidgets<8.0,>=7.5->dataprep) (57.2.0)

Requirement already satisfied: pygments in /usr/local/lib/python3.7/dist-packages (from ipython>=4.0.0->ipywidgets<8.0,>=7.5->dataprep) (2.6.1)

Requirement already satisfied: simplegeneric>0.8 in /usr/local/lib/python3.7/dist-packages (from ipython>=4.0.0->ipywidgets<8.0,>=7.5->dataprep) (0.8.1)

Requirement already satisfied: pickleshare in /usr/local/lib/python3.7/dist-packages (from ipython>=4.0.0->ipywidgets<8.0,>=7.5->dataprep) (0.7.5)

Requirement already satisfied: decorator in /usr/local/lib/python3.7/dist-packages (from ipython>=4.0.0->ipywidgets<8.0,>=7.5->dataprep) (4.4.2)

Requirement already satisfied: prompt-toolkit<2.0.0,>=1.0.4 in /usr/local/lib/python3.7/dist-packages (from ipython>=4.0.0->ipywidgets<8.0,>=7.5->dataprep) (1.0.18)

Requirement already satisfied: MarkupSafe>=0.23 in /usr/local/lib/python3.7/dist-packages (from jinja2<3.0,>=2.11->dataprep) (2.0.1)

Collecting ply

Downloading ply-3.11-py2.py3-none-any.whl (49 kB)

|████████████████████████████████| 49 kB 5.0 MB/s

Requirement already satisfied: six in /usr/local/lib/python3.7/dist-packages (from jsonpath-ng<2.0,>=1.5->dataprep) (1.15.0)

Requirement already satisfied: ipython-genutils in /usr/local/lib/python3.7/dist-packages (from nbformat>=4.2.0->ipywidgets<8.0,>=7.5->dataprep) (0.2.0)

Requirement already satisfied: jupyter-core in /usr/local/lib/python3.7/dist-packages (from nbformat>=4.2.0->ipywidgets<8.0,>=7.5->dataprep) (4.7.1)

Requirement already satisfied: jsonschema!=2.5.0,>=2.4 in /usr/local/lib/python3.7/dist-packages (from nbformat>=4.2.0->ipywidgets<8.0,>=7.5->dataprep) (2.6.0)

Requirement already satisfied: click in /usr/local/lib/python3.7/dist-packages (from nltk<4.0,>=3.5->dataprep) (7.1.2)

Requirement already satisfied: joblib in /usr/local/lib/python3.7/dist-packages (from nltk<4.0,>=3.5->dataprep) (1.0.1)

Requirement already satisfied: pyparsing>=2.0.2 in /usr/local/lib/python3.7/dist-packages (from packaging>=16.8->bokeh<3,>=2->dataprep) (2.4.7)

Requirement already satisfied: pytz>=2017.2 in /usr/local/lib/python3.7/dist-packages (from pandas<2.0,>=1.1->dataprep) (2018.9)

Collecting locket

Downloading locket-0.2.1-py2.py3-none-any.whl (4.1 kB)

Requirement already satisfied: wcwidth in /usr/local/lib/python3.7/dist-packages (from prompt-toolkit<2.0.0,>=1.0.4->ipython>=4.0.0->ipywidgets<8.0,>=7.5->dataprep) (0.2.5)

Collecting python-crfsuite>=0.7

Downloading python\_crfsuite-0.9.7-cp37-cp37m-manylinux1\_x86\_64.whl (743 kB)

|████████████████████████████████| 743 kB 39.9 MB/s

Requirement already satisfied: future>=0.14 in /usr/local/lib/python3.7/dist-packages (from usaddress<0.6.0,>=0.5.10->dataprep) (0.16.0)

Collecting probableparsing

Downloading probableparsing-0.0.1-py2.py3-none-any.whl (3.1 kB)

Requirement already satisfied: notebook>=4.4.1 in /usr/local/lib/python3.7/dist-packages (from widgetsnbextension~=3.5.0->ipywidgets<8.0,>=7.5->dataprep) (5.3.1)

Requirement already satisfied: terminado>=0.8.1 in /usr/local/lib/python3.7/dist-packages (from notebook>=4.4.1->widgetsnbextension~=3.5.0->ipywidgets<8.0,>=7.5->dataprep) (0.10.1)

Requirement already satisfied: Send2Trash in /usr/local/lib/python3.7/dist-packages (from notebook>=4.4.1->widgetsnbextension~=3.5.0->ipywidgets<8.0,>=7.5->dataprep) (1.7.1)

Requirement already satisfied: nbconvert in /usr/local/lib/python3.7/dist-packages (from notebook>=4.4.1->widgetsnbextension~=3.5.0->ipywidgets<8.0,>=7.5->dataprep) (5.6.1)

Requirement already satisfied: pyzmq>=13 in /usr/local/lib/python3.7/dist-packages (from jupyter-client->ipykernel>=4.5.1->ipywidgets<8.0,>=7.5->dataprep) (22.1.0)

Requirement already satisfied: ptyprocess in /usr/local/lib/python3.7/dist-packages (from terminado>=0.8.1->notebook>=4.4.1->widgetsnbextension~=3.5.0->ipywidgets<8.0,>=7.5->dataprep) (0.7.0)

Requirement already satisfied: matplotlib in /usr/local/lib/python3.7/dist-packages (from wordcloud<2.0,>=1.8->dataprep) (3.2.2)

Requirement already satisfied: idna>=2.0 in /usr/local/lib/python3.7/dist-packages (from yarl<2.0,>=1.0->aiohttp<4.0,>=3.6->dataprep) (2.10)

Requirement already satisfied: kiwisolver>=1.0.1 in /usr/local/lib/python3.7/dist-packages (from matplotlib->wordcloud<2.0,>=1.8->dataprep) (1.3.1)

Requirement already satisfied: cycler>=0.10 in /usr/local/lib/python3.7/dist-packages (from matplotlib->wordcloud<2.0,>=1.8->dataprep) (0.10.0)

Requirement already satisfied: testpath in /usr/local/lib/python3.7/dist-packages (from nbconvert->notebook>=4.4.1->widgetsnbextension~=3.5.0->ipywidgets<8.0,>=7.5->dataprep) (0.5.0)

Requirement already satisfied: mistune<2,>=0.8.1 in /usr/local/lib/python3.7/dist-packages (from nbconvert->notebook>=4.4.1->widgetsnbextension~=3.5.0->ipywidgets<8.0,>=7.5->dataprep) (0.8.4)

Requirement already satisfied: pandocfilters>=1.4.1 in /usr/local/lib/python3.7/dist-packages (from nbconvert->notebook>=4.4.1->widgetsnbextension~=3.5.0->ipywidgets<8.0,>=7.5->dataprep) (1.4.3)

Requirement already satisfied: entrypoints>=0.2.2 in /usr/local/lib/python3.7/dist-packages (from nbconvert->notebook>=4.4.1->widgetsnbextension~=3.5.0->ipywidgets<8.0,>=7.5->dataprep) (0.3)

Requirement already satisfied: bleach in /usr/local/lib/python3.7/dist-packages (from nbconvert->notebook>=4.4.1->widgetsnbextension~=3.5.0->ipywidgets<8.0,>=7.5->dataprep) (3.3.0)

Requirement already satisfied: defusedxml in /usr/local/lib/python3.7/dist-packages (from nbconvert->notebook>=4.4.1->widgetsnbextension~=3.5.0->ipywidgets<8.0,>=7.5->dataprep) (0.7.1)

Requirement already satisfied: webencodings in /usr/local/lib/python3.7/dist-packages (from bleach->nbconvert->notebook>=4.4.1->widgetsnbextension~=3.5.0->ipywidgets<8.0,>=7.5->dataprep) (0.5.1)

Building wheels for collected packages: metaphone, python-Levenshtein

Building wheel for metaphone (setup.py) ... done

Created wheel for metaphone: filename=Metaphone-0.6-py3-none-any.whl size=13919 sha256=215d20a6cdbce27c897f13386dfb9e2cd6694c02eaa3751f116ab06fb02f7e2f

Stored in directory: /root/.cache/pip/wheels/1d/a8/cb/6f8902aa5457bd71344e00665c230e9c45255b3f57f2194a0f

Building wheel for python-Levenshtein (setup.py) ... done

Created wheel for python-Levenshtein: filename=python\_Levenshtein-0.12.2-cp37-cp37m-linux\_x86\_64.whl size=149859 sha256=bee4ad707bc9163790fb8367d302ec93009a58b21ea866298df2a84bf883249c

Stored in directory: /root/.cache/pip/wheels/05/5f/ca/7c4367734892581bb5ff896f15027a932c551080b2abd3e00d

Successfully built metaphone python-Levenshtein

Installing collected packages: multidict, locket, yarl, tqdm, regex, python-crfsuite, probableparsing, ply, partd, fsspec, dask, async-timeout, wordcloud, usaddress, python-Levenshtein, pydantic, nltk, metaphone, jsonpath-ng, aiohttp, dataprep

Attempting uninstall: tqdm

Found existing installation: tqdm 4.41.1

Uninstalling tqdm-4.41.1:

Successfully uninstalled tqdm-4.41.1

Attempting uninstall: regex

Found existing installation: regex 2019.12.20

Uninstalling regex-2019.12.20:

Successfully uninstalled regex-2019.12.20

Attempting uninstall: dask

Found existing installation: dask 2.12.0

Uninstalling dask-2.12.0:

Successfully uninstalled dask-2.12.0

Attempting uninstall: wordcloud

Found existing installation: wordcloud 1.5.0

Uninstalling wordcloud-1.5.0:

Successfully uninstalled wordcloud-1.5.0

Attempting uninstall: nltk

Found existing installation: nltk 3.2.5

Uninstalling nltk-3.2.5:

Successfully uninstalled nltk-3.2.5

Successfully installed aiohttp-3.7.4.post0 async-timeout-3.0.1 dask-2.30.0 dataprep-0.3.0 fsspec-2021.7.0 jsonpath-ng-1.5.3 locket-0.2.1 metaphone-0.6 multidict-5.1.0 nltk-3.6.2 partd-1.2.0 ply-3.11 probableparsing-0.0.1 pydantic-1.8.2 python-Levenshtein-0.12.2 python-crfsuite-0.9.7 regex-2020.11.13 tqdm-4.62.0 usaddress-0.5.10 wordcloud-1.8.1 yarl-1.6.3

[ ]

!apt-get install openjdk-8-jdk-headless -qq > /dev/null

!wget -q https://archive.apache.org/dist/spark/spark-2.4.4/spark-2.4.4-bin-hadoop2.7.tgz

!tar xf spark-2.4.4-bin-hadoop2.7.tgz

!pip install -q findspark

# configurar as variáveis de ambiente

import os

os.environ["JAVA\_HOME"] = "/usr/lib/jvm/java-8-openjdk-amd64"

os.environ["SPARK\_HOME"] = "/content/spark-2.4.4-bin-hadoop2.7"

# tornar o pyspark "importável"

import findspark

findspark.init('spark-2.4.4-bin-hadoop2.7')





[ ]

from pyspark.sql import SparkSession

sc = SparkSession.builder.master('local[\*]').getOrCreate()

df = "/content/clientes.csv"

clientes = sc.read.format("csv").options(inferSchema=True, header=True, sep=',').load(df).toPandas()





[ ]

from dataprep.eda import create\_report

create\_report(clientes)





[ ]

#Verifica as Dimensões do Arquivo

clientes.shape





(614, 13)

[ ]

#Verificando os primeiros registros

clientes.head()





[ ]

#Verificando as informações dos campos da tabela de dados

clientes.info()





<class 'pandas.core.frame.DataFrame'>

RangeIndex: 614 entries, 0 to 613

Data columns (total 13 columns):

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 cod\_cliente 614 non-null object

1 sexo 601 non-null object

2 estado\_civil 611 non-null object

3 dependentes 599 non-null object

4 educacao 614 non-null object

5 empregado 582 non-null object

6 renda 614 non-null int32

7 renda\_conjuge 614 non-null object

8 emprestimo 592 non-null float64

9 prestacao\_mensal 600 non-null float64

10 historico\_credito 564 non-null float64

11 imovel 614 non-null object

12 aprovacao\_emprestimo 614 non-null object

dtypes: float64(3), int32(1), object(9)

memory usage: 60.1+ KB

[ ]

#Verifica se há valores Nulos/ausentes

clientes.isnull().sum()





cod\_cliente 0

sexo 13

estado\_civil 3

dependentes 15

educacao 0

empregado 32

renda 0

renda\_conjuge 0

emprestimo 22

prestacao\_mensal 14

historico\_credito 50

imovel 0

aprovacao\_emprestimo 0

dtype: int64

[ ]

#verifica se há valores duplicados

clientes.duplicated().sum()





0

Conhecendo a Distribuição dos Dados de Cada Variável/Campo

[ ]

#Estatísticas descritivas das variáveis numéricas

clientes.describe()





[ ]

# Tabela de Frequência da Variável "aprovacao\_emprestimo" - Nossa Classe ou Label ou Target ou Y ou Variável a ser Predita (o que a gente quer descobrir)

clientes.aprovacao\_emprestimo.value\_counts()





Y 422

N 192

Name: aprovacao\_emprestimo, dtype: int64

[ ]

#Criando Gráfico de Barras para variáveis Categóricas

sns.countplot(clientes['aprovacao\_emprestimo'])





[ ]

#Criando Gráfico de Barras para variáveis Categóricas

sns.countplot(clientes['historico\_credito'])





[ ]

#Criando Gráfico de Barras para variáveis Categóricas

sns.countplot(clientes['educacao'])





[ ]

#Criando Gráfico de Barras para variáveis Categóricas

sns.countplot(clientes['sexo'])





[ ]

#Criando Gráfico de Barras para variáveis Categóricas

sns.countplot(clientes['imovel'])





[ ]

#Criando Gráfico de Barras para variáveis Categóricas

sns.countplot(clientes['educacao'])





[ ]

#Criando Gráfico de Barras para variáveis Categóricas

sns.countplot(clientes['dependentes'])





[ ]

#Criando Gráfico de Barras para variáveis Categóricas

sns.countplot(clientes['estado\_civil'])





[ ]

#Criando Gráfico de Barras para variáveis Categóricas

sns.countplot(clientes['empregado'])





[ ]

#Criando Gráfico de Distribuição para variáveis contínuas

sns.histplot(clientes['renda'], bins=50, kde=True)





[ ]

#Distribuição dos dados BOXPLOT

clientes.boxplot(column='emprestimo', vert=False )





[ ]

#Criando Gráfico de Distribuição para variáveis contínuas

sns.histplot(clientes['emprestimo'], bins=50, kde=True)





[ ]

# Correlação entre as Variáveis

sns.scatterplot(data=clientes, x="emprestimo", y="renda", ) #hue='aprovacao\_emprestimo')





[ ]

# Criação do Gráfico de dispersão (scatter) conhecido como pairplot

sns.pairplot(clientes, hue='aprovacao\_emprestimo') # kind=reg é a reta de regressão





[ ]

# Eliminando valores Nulos

clientes = clientes.dropna(subset=['sexo','estado\_civil','renda','emprestimo','historico\_credito'])

clientes.isnull().sum()





cod\_cliente 0

sexo 0

estado\_civil 0

dependentes 11

educacao 0

empregado 26

renda 0

renda\_conjuge 0

emprestimo 0

prestacao\_mensal 14

historico\_credito 0

imovel 0

aprovacao\_emprestimo 0

dtype: int64

[ ]

# Buscando ferramentas para gerar vetores das Strings e criando dataframe para leitura no spark

from pyspark.ml.feature import VectorAssembler,OneHotEncoder, StringIndexer

from pyspark.ml import Pipeline

sparkdata=sc.createDataFrame(clientes)

sparkdata.show()





+-----------+------+------------+-----------+------------+---------+-----+-------------+----------+----------------+-----------------+---------+--------------------+

|cod\_cliente| sexo|estado\_civil|dependentes| educacao|empregado|renda|renda\_conjuge|emprestimo|prestacao\_mensal|historico\_credito| imovel|aprovacao\_emprestimo|

+-----------+------+------------+-----------+------------+---------+-----+-------------+----------+----------------+-----------------+---------+--------------------+

| LP001003| Male| Yes| 1| Graduate| No| 4583| 1508| 128.0| 360.0| 1.0| Rural| N|

| LP001005| Male| Yes| 0| Graduate| Yes| 3000| 0| 66.0| 360.0| 1.0| Urban| Y|

| LP001006| Male| Yes| 0|Not Graduate| No| 2583| 2358| 120.0| 360.0| 1.0| Urban| Y|

| LP001008| Male| No| 0| Graduate| No| 6000| 0| 141.0| 360.0| 1.0| Urban| Y|

| LP001011| Male| Yes| 2| Graduate| Yes| 5417| 4196| 267.0| 360.0| 1.0| Urban| Y|

| LP001013| Male| Yes| 0|Not Graduate| No| 2333| 1516| 95.0| 360.0| 1.0| Urban| Y|

| LP001014| Male| Yes| 3+| Graduate| No| 3036| 2504| 158.0| 360.0| 0.0|Semiurban| N|

| LP001018| Male| Yes| 2| Graduate| No| 4006| 1526| 168.0| 360.0| 1.0| Urban| Y|

| LP001020| Male| Yes| 1| Graduate| No|12841| 10968| 349.0| 360.0| 1.0|Semiurban| N|

| LP001024| Male| Yes| 2| Graduate| No| 3200| 700| 70.0| 360.0| 1.0| Urban| Y|

| LP001027| Male| Yes| 2| Graduate| null| 2500| 1840| 109.0| 360.0| 1.0| Urban| Y|

| LP001028| Male| Yes| 2| Graduate| No| 3073| 8106| 200.0| 360.0| 1.0| Urban| Y|

| LP001029| Male| No| 0| Graduate| No| 1853| 2840| 114.0| 360.0| 1.0| Rural| N|

| LP001030| Male| Yes| 2| Graduate| No| 1299| 1086| 17.0| 120.0| 1.0| Urban| Y|

| LP001032| Male| No| 0| Graduate| No| 4950| 0| 125.0| 360.0| 1.0| Urban| Y|

| LP001036|Female| No| 0| Graduate| No| 3510| 0| 76.0| 360.0| 0.0| Urban| N|

| LP001038| Male| Yes| 0|Not Graduate| No| 4887| 0| 133.0| 360.0| 1.0| Rural| N|

| LP001041| Male| Yes| 0| Graduate| null| 2600| 3500| 115.0| NaN| 1.0| Urban| Y|

| LP001043| Male| Yes| 0|Not Graduate| No| 7660| 0| 104.0| 360.0| 0.0| Urban| N|

| LP001046| Male| Yes| 1| Graduate| No| 5955| 5625| 315.0| 360.0| 1.0| Urban| Y|

+-----------+------+------------+-----------+------------+---------+-----+-------------+----------+----------------+-----------------+---------+--------------------+

only showing top 20 rows

[ ]

stringIndexer\_sexo=StringIndexer(inputCol="sexo", outputCol="sexo\_encoded")  #label encoding

encoder\_sexo = OneHotEncoder(dropLast=False, inputCol="sexo\_encoded", outputCol="sexoVec") #one-hot encoding

stringIndexer\_estado\_civil=StringIndexer(inputCol="estado\_civil", outputCol="estado\_civil\_encoded")  #label encoding

encoder\_estado\_civil = OneHotEncoder(dropLast=False, inputCol="estado\_civil\_encoded", outputCol="estado\_civilVec") #one-hot encoding

stringIndexer\_aprovacao\_emprestimo=StringIndexer(inputCol="aprovacao\_emprestimo", outputCol="aprovacao\_emprestimo\_encoded")  #label encoding

encoder\_aprovacao\_emprestimo = OneHotEncoder(dropLast=False, inputCol="aprovacao\_emprestimo\_encoded", outputCol="aprovacao\_emprestimoVec") #one-hot encoding

colunas\_entrada=['sexoVec','estado\_civilVec','renda','emprestimo','historico\_credito']

vetor\_entrada = VectorAssembler(inputCols=colunas\_entrada,outputCol='features')

sequencia\_transformacoes=[stringIndexer\_sexo,stringIndexer\_estado\_civil,stringIndexer\_aprovacao\_emprestimo,encoder\_sexo,encoder\_estado\_civil,encoder\_aprovacao\_emprestimo,vetor\_entrada]





[ ]

# Aplicando o pipeline

!pip install -q lib5c

!lib5c -v

from pyspark.ml import Pipeline





|████████████████████████████████| 16.2 MB 154 kB/s

|████████████████████████████████| 1.2 MB 43.3 MB/s

Building wheel for interlap (setup.py) ... done

Building wheel for luigi (setup.py) ... done

Building wheel for powerlaw (setup.py) ... done

lib5c version 0.6.1

[ ]

pipeline = Pipeline(stages=sequencia\_transformacoes)

pipelineModel = pipeline.fit(sparkdata)

model = pipelineModel.transform(sparkdata)





[ ]

model.show()





+-----------+------+------------+-----------+------------+---------+-----+-------------+----------+----------------+-----------------+---------+--------------------+------------+--------------------+----------------------------+-------------+---------------+-----------------------+--------------------+

|cod\_cliente| sexo|estado\_civil|dependentes| educacao|empregado|renda|renda\_conjuge|emprestimo|prestacao\_mensal|historico\_credito| imovel|aprovacao\_emprestimo|sexo\_encoded|estado\_civil\_encoded|aprovacao\_emprestimo\_encoded| sexoVec|estado\_civilVec|aprovacao\_emprestimoVec| features|

+-----------+------+------------+-----------+------------+---------+-----+-------------+----------+----------------+-----------------+---------+--------------------+------------+--------------------+----------------------------+-------------+---------------+-----------------------+--------------------+

| LP001003| Male| Yes| 1| Graduate| No| 4583| 1508| 128.0| 360.0| 1.0| Rural| N| 0.0| 0.0| 1.0|(2,[0],[1.0])| (2,[0],[1.0])| (2,[1],[1.0])|[1.0,0.0,1.0,0.0,...|

| LP001005| Male| Yes| 0| Graduate| Yes| 3000| 0| 66.0| 360.0| 1.0| Urban| Y| 0.0| 0.0| 0.0|(2,[0],[1.0])| (2,[0],[1.0])| (2,[0],[1.0])|[1.0,0.0,1.0,0.0,...|

| LP001006| Male| Yes| 0|Not Graduate| No| 2583| 2358| 120.0| 360.0| 1.0| Urban| Y| 0.0| 0.0| 0.0|(2,[0],[1.0])| (2,[0],[1.0])| (2,[0],[1.0])|[1.0,0.0,1.0,0.0,...|

| LP001008| Male| No| 0| Graduate| No| 6000| 0| 141.0| 360.0| 1.0| Urban| Y| 0.0| 1.0| 0.0|(2,[0],[1.0])| (2,[1],[1.0])| (2,[0],[1.0])|[1.0,0.0,0.0,1.0,...|

| LP001011| Male| Yes| 2| Graduate| Yes| 5417| 4196| 267.0| 360.0| 1.0| Urban| Y| 0.0| 0.0| 0.0|(2,[0],[1.0])| (2,[0],[1.0])| (2,[0],[1.0])|[1.0,0.0,1.0,0.0,...|

| LP001013| Male| Yes| 0|Not Graduate| No| 2333| 1516| 95.0| 360.0| 1.0| Urban| Y| 0.0| 0.0| 0.0|(2,[0],[1.0])| (2,[0],[1.0])| (2,[0],[1.0])|[1.0,0.0,1.0,0.0,...|

| LP001014| Male| Yes| 3+| Graduate| No| 3036| 2504| 158.0| 360.0| 0.0|Semiurban| N| 0.0| 0.0| 1.0|(2,[0],[1.0])| (2,[0],[1.0])| (2,[1],[1.0])|[1.0,0.0,1.0,0.0,...|

| LP001018| Male| Yes| 2| Graduate| No| 4006| 1526| 168.0| 360.0| 1.0| Urban| Y| 0.0| 0.0| 0.0|(2,[0],[1.0])| (2,[0],[1.0])| (2,[0],[1.0])|[1.0,0.0,1.0,0.0,...|

| LP001020| Male| Yes| 1| Graduate| No|12841| 10968| 349.0| 360.0| 1.0|Semiurban| N| 0.0| 0.0| 1.0|(2,[0],[1.0])| (2,[0],[1.0])| (2,[1],[1.0])|[1.0,0.0,1.0,0.0,...|

| LP001024| Male| Yes| 2| Graduate| No| 3200| 700| 70.0| 360.0| 1.0| Urban| Y| 0.0| 0.0| 0.0|(2,[0],[1.0])| (2,[0],[1.0])| (2,[0],[1.0])|[1.0,0.0,1.0,0.0,...|

| LP001028| Male| Yes| 2| Graduate| No| 3073| 8106| 200.0| 360.0| 1.0| Urban| Y| 0.0| 0.0| 0.0|(2,[0],[1.0])| (2,[0],[1.0])| (2,[0],[1.0])|[1.0,0.0,1.0,0.0,...|

| LP001029| Male| No| 0| Graduate| No| 1853| 2840| 114.0| 360.0| 1.0| Rural| N| 0.0| 1.0| 1.0|(2,[0],[1.0])| (2,[1],[1.0])| (2,[1],[1.0])|[1.0,0.0,0.0,1.0,...|

| LP001030| Male| Yes| 2| Graduate| No| 1299| 1086| 17.0| 120.0| 1.0| Urban| Y| 0.0| 0.0| 0.0|(2,[0],[1.0])| (2,[0],[1.0])| (2,[0],[1.0])|[1.0,0.0,1.0,0.0,...|

| LP001032| Male| No| 0| Graduate| No| 4950| 0| 125.0| 360.0| 1.0| Urban| Y| 0.0| 1.0| 0.0|(2,[0],[1.0])| (2,[1],[1.0])| (2,[0],[1.0])|[1.0,0.0,0.0,1.0,...|

| LP001036|Female| No| 0| Graduate| No| 3510| 0| 76.0| 360.0| 0.0| Urban| N| 1.0| 1.0| 1.0|(2,[1],[1.0])| (2,[1],[1.0])| (2,[1],[1.0])|[0.0,1.0,0.0,1.0,...|

| LP001038| Male| Yes| 0|Not Graduate| No| 4887| 0| 133.0| 360.0| 1.0| Rural| N| 0.0| 0.0| 1.0|(2,[0],[1.0])| (2,[0],[1.0])| (2,[1],[1.0])|[1.0,0.0,1.0,0.0,...|

| LP001043| Male| Yes| 0|Not Graduate| No| 7660| 0| 104.0| 360.0| 0.0| Urban| N| 0.0| 0.0| 1.0|(2,[0],[1.0])| (2,[0],[1.0])| (2,[1],[1.0])|[1.0,0.0,1.0,0.0,...|

| LP001046| Male| Yes| 1| Graduate| No| 5955| 5625| 315.0| 360.0| 1.0| Urban| Y| 0.0| 0.0| 0.0|(2,[0],[1.0])| (2,[0],[1.0])| (2,[0],[1.0])|[1.0,0.0,1.0,0.0,...|

| LP001047| Male| Yes| 0|Not Graduate| No| 2600| 1911| 116.0| 360.0| 0.0|Semiurban| N| 0.0| 0.0| 1.0|(2,[0],[1.0])| (2,[0],[1.0])| (2,[1],[1.0])|[1.0,0.0,1.0,0.0,...|

| LP001066| Male| Yes| 0| Graduate| Yes| 9560| 0| 191.0| 360.0| 1.0|Semiurban| Y| 0.0| 0.0| 0.0|(2,[0],[1.0])| (2,[0],[1.0])| (2,[0],[1.0])|[1.0,0.0,1.0,0.0,...|

+-----------+------+------------+-----------+------------+---------+-----+-------------+----------+----------------+-----------------+---------+--------------------+------------+--------------------+----------------------------+-------------+---------------+-----------------------+--------------------+

only showing top 20 rows

[ ]

#divide os dados para treino e testes

train\_data, test\_data = model.randomSplit([.8,.2],seed=5)

#define o modelo de regrssão logística

from pyspark.ml.classification import LogisticRegression

#instancia o objeto para a regressão logística

lr = LogisticRegression(labelCol="aprovacao\_emprestimo\_encoded",featuresCol="features", maxIter=100, regParam=0.3)

# treina o modelo

linearModel = lr.fit(train\_data)





[ ]

#avaliação da acuracia do modelo

previsao\_regressao = linearModel.transform(test\_data)

from pyspark.ml.evaluation import MulticlassClassificationEvaluator

acc\_evaluator = MulticlassClassificationEvaluator(labelCol="aprovacao\_emprestimo\_encoded", predictionCol="prediction", metricName="accuracy")

acuracia\_regressao = acc\_evaluator.evaluate(previsao\_regressao)

print('Regressão Logística: {0:2.2f}%'.format(acuracia\_regressao\*100))





Regressão Logística: 82.73%

[ ]

#\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_





[ ]

teste2 = clientes.copy()

teste2





[ ]

teste2['sexo']= teste2['sexo'].map({'Male':0, 'Female':1})

teste2['estado\_civil']= teste2['estado\_civil'].map({'No':0, 'Yes':1})

teste2['aprovacao\_emprestimo']= teste2['aprovacao\_emprestimo'].map({'N':0, 'Y':1})





[ ]

teste2





[ ]

X = teste2[['sexo', 'estado\_civil', 'renda', 'emprestimo', 'historico\_credito']]

y = teste2.aprovacao\_emprestimo

X.shape, y.shape





((529, 5), (529,))

[ ]

from sklearn.model\_selection import train\_test\_split





[ ]

x\_train, x\_teste, y\_train, y\_teste = train\_test\_split(X,y, test\_size = 0.2, random\_state = 7)





[ ]

from sklearn.ensemble import RandomForestClassifier





[ ]

maquina = RandomForestClassifier(max\_depth=4, random\_state=7)

maquina.fit(x\_train,y\_train)





RandomForestClassifier(bootstrap=True, ccp\_alpha=0.0, class\_weight=None,

criterion='gini', max\_depth=4, max\_features='auto',

max\_leaf\_nodes=None, max\_samples=None,

min\_impurity\_decrease=0.0, min\_impurity\_split=None,

min\_samples\_leaf=1, min\_samples\_split=2,

min\_weight\_fraction\_leaf=0.0, n\_estimators=100,

n\_jobs=None, oob\_score=False, random\_state=2, verbose=0,

warm\_start=False)

[ ]

from sklearn.metrics import accuracy\_score





[ ]

pred\_maquina = maquina.predict(x\_train)

accuracy\_score(y\_train,pred\_maquina)





0.8132387706855791

[ ]

pred\_maquina = maquina.predict(x\_teste)

accuracy\_score(y\_teste,pred\_maquina)





[ ]

maquina.feature\_importances\_



