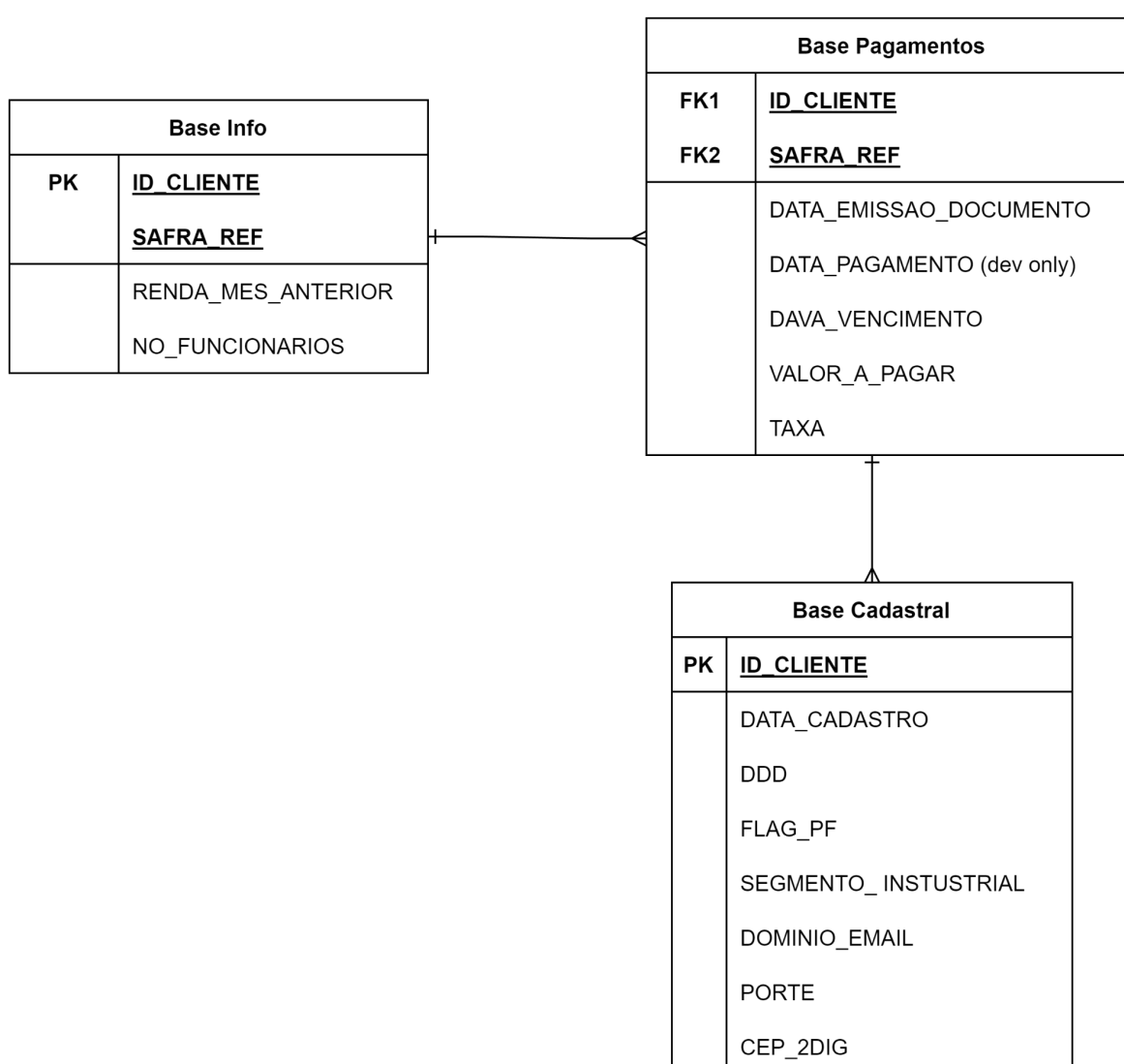


## Instruções para o case

Modelos de score de crédito calculam a probabilidade de inadimplência e são uma das principais ferramentas utilizadas por diversas empresas para aprovar ou negar um crédito.

O objetivo deste desafio é criar um modelo preditivo calculando a probabilidade de inadimplência de cada novo pedido de crédito realizado por um cliente recorrente.

Para o desafio, são disponibilizadas 3 bases de dados contendo informações de cada cliente e transação realizada por ele. O relacionamento entre essas bases é como segue.



# Dados

## Base Cadastral

Arquivo: base\_cadastral.csv

Base contendo informações cadastrais dos clientes. Cada cliente deve ter apenas uma data de cadastro e seus dados não mudam ao longo do tempo.

Variáveis:

- ID\_CLIENTE: Identificador único do cliente.
- DATA\_CADASTRO: Data da realização do cadastro no sistema.
- DDD: Número do DDD do telefone do cliente.
- FLAG\_PF: Indica se o cliente é uma pessoa física ('X') ou jurídica ('NaN').
- SEGMENTO\_INDUSTRIAL: Indica a qual segmento da indústria pertence o cliente.
- DOMINIO\_EMAIL: Indica o domínio(ou provedor) do email utilizado para o cadastro.
- PORTE: Indica o porte (tamanho) da empresa.
- CEP\_2\_DIG: Indica os dois primeiros números do CEP do endereço cadastrado.

## Base Info

Arquivo: base\_info.csv

Base com informações adicionais dos clientes. Essa base é atualizada mensalmente, então cada cliente aparecerá apenas uma vez em cada mês de referência. Ou seja, o identificador único da base consiste na combinação do ID\_CLIENTE e da SAFRA\_REF.

Variáveis:

- ID\_CLIENTE: Identificador único do cliente.
- SAFRA\_REF: Mês de referência da amostra.
- RENDA\_MES\_ANTERIOR: Renda ou faturamento declarado pelo cliente no fim do mês anterior.
- NO\_FUNCIONARIOS: Número de funcionários reportado pelo cliente no fim do mês anterior.

## Base Pagamentos

Arquivos:

- base\_pagamentos\_desenvolvimento.csv (desenvolvimento do modelo)
- base\_pagamentos\_teste.csv (validação interna)

Base com informações sobre transações (empréstimos) passados. Cada cliente pode ter uma ou mais transações no mesmo período de tempo.

- ID\_CLIENTE: Identificador único do cliente.
- SAFRA\_REF: Mês de referência da amostra.
- DATA\_EMISSAO\_DOCUMENTO: Data da emissão da nota de crédito.
- DATA\_VENCIMENTO: Data limite para pagamento do empréstimo.
- VALOR\_A\_PAGAR: Valor da nota de crédito.
- TAXA: Taxa de juros cobrada no empréstimo.
- DATA\_PAGAMENTO: Data em que o cliente realizou o pagamento da nota (disponível apenas na base de desenvolvimento).

## Variável resposta (target)

A variável resposta do modelo será a inadimplência do cliente em determinada transação.

Para simplificar sua construção, serão considerados inadimplentes os clientes que possuem atraso, ou seja, diferença entre o pagamento e o vencimento, maior ou igual a 5 dias.

**Obs.:** Os dados apresentados nas bases são anonimizados e possuem parte sintética, entretanto, é possível que existam inconsistências e incoerências comuns em dados reais. Cabe ao candidato tratar os erros encontrados nos dados da forma que considerar mais adequada.

## Modelo

Utilizando os dados fornecidos, desenvolva um modelo preditivo capaz de gerar previsões a respeito das amostras presentes na base base\_pagamentos\_teste.csv. Você deve gerar uma base contendo as colunas: **ID\_CLIENTE, SAFRA\_REF e INADIMPLENTE**, sendo a última a probabilidade do cliente ser inadimplente.

## Avaliação

Serão avaliados o código, preferencialmente em python ou R, e o desempenho do modelo.

Para isto, será necessário enviar o código com tudo o que foi feito, desde análises exploratórias até a previsão dos dados de teste/validação e a base de dados de teste com a coluna de previsões.