UFMG
UNIVERSIDADE FEDERAL
DE MINAS GERAIS

EXERCÍCIO
REGRESSÃO LOGÍSTICA

ALUNO: FELIPPE VELOSO MARINHO
MATRÍCULA: 2021072260
DISCIPLINA: APRENDIZADO DE MÁQUINA

Objetivo: Análise de Inadimplência no Conjunto de Dados *Credit Card Default* com *Regressão Logística*.

1. Implementar uma regressão logística para prever a inadimplência (default) com base nas variáveis balance, student, income e outras variáveis relevantes.
2. Analisar os coeficientes do modelo e a significância estatística das variáveis.
3. Avaliar o desempenho do modelo usando métricas de classificação e interpretação dos resultados

Após carregar o conjunto de dados **Credit Card Default Data Set.** Nele podemos conferir um conjunto de dados que busca nos dizer se o cliente está inadimplente ou não, relacionando com o fato dele ser estudante. A variável ***default*** é o nosso **target** do problema. Ela indica se o cliente está inadimplente (1) ou não (0).

A variável **balance** nos indica o **saldo médio do cliente** no cartão de crédito.

O **income** é a **renda anual** do cliente.

	target	estudante	saldo médio	renda anual
0	No	No	729.526495	44361.625074
1	No	Yes	817.180407	12106.134700
2	No	No	1073.549164	31767.138947
3	No	No	529.250605	35704.493935
4	No	No	785.655883	38463.495879

Realizamos a verificação do resumo estatístico e a distribuição das variáveis principais na intenção de encontrar algumas relações. Em primeiro ponto, é possível

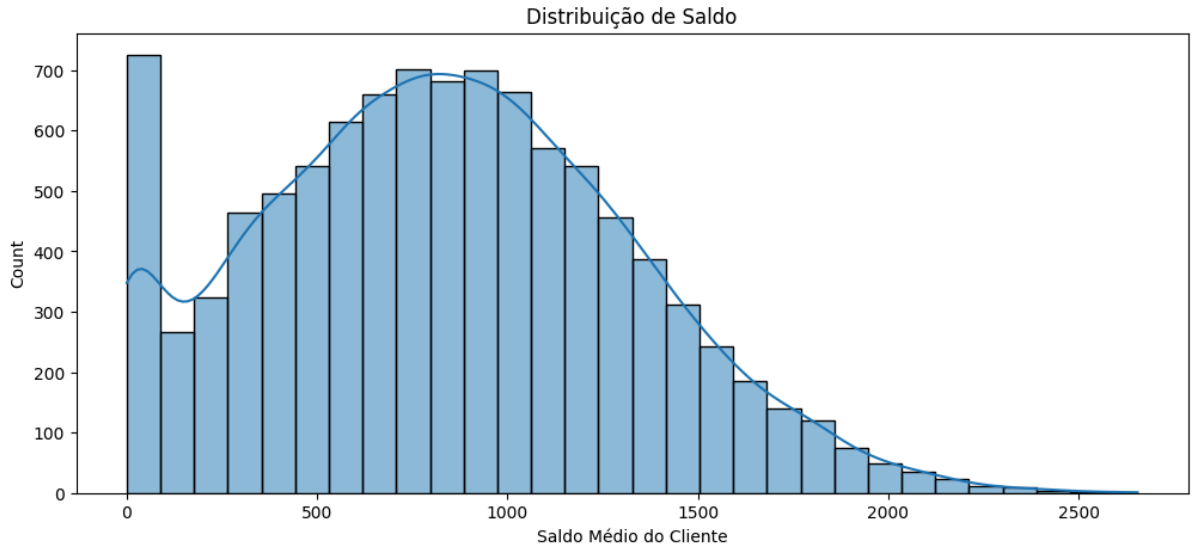
verificar que o desvio padrão de saldo médio é de `483.71`, indicando uma grande variação nos saldos, sugerindo que existem clientes com saldos bem diferentes entre si. Já o de renda anual é de 13336,64 indica uma variação bem ampla entre os clientes.

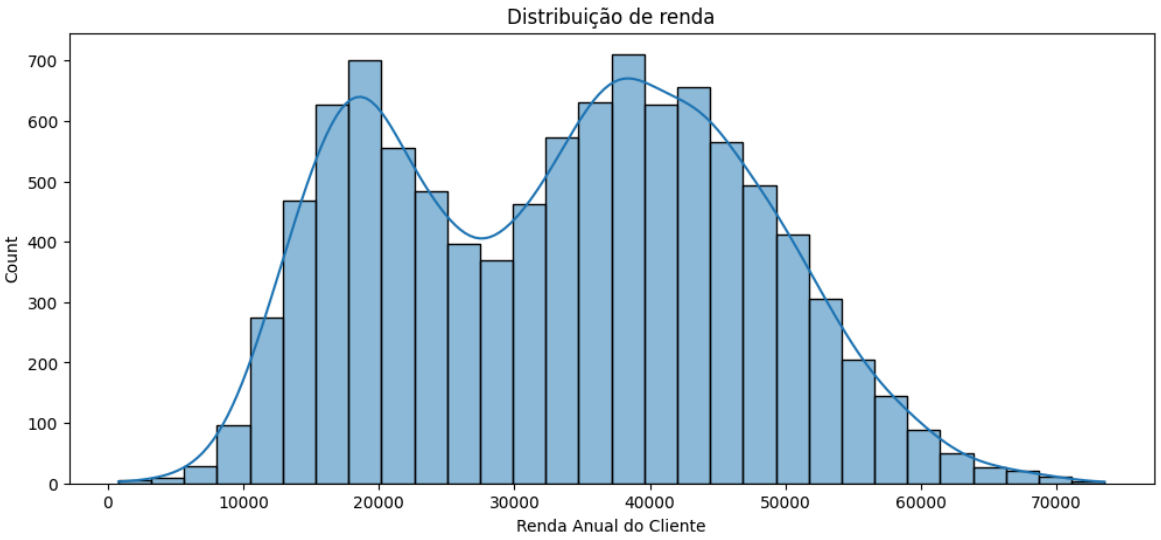
```
Valores ausentes em cada coluna:
target         0
estudante      0
saldo médio    0
renda anual    0
dtype: int64

Tipos de dados:
target         category
estudante      category
saldo médio     float64
renda anual     float64
dtype: object

Distribuição das variáveis categóricas:
estudante
No     7056
Yes    2944
Name: count, dtype: int64
```

	saldo médio	renda anual
count	10000.000000	10000.000000
mean	835.374886	33516.981876
std	483.714985	13336.639563
min	0.000000	771.967729
25%	481.731105	21340.462903
50%	823.636973	34552.644802
75%	1166.308386	43807.729272
max	2654.322576	73554.233495

O alto desvio padrão e a diferença entre o valor médio e o máximo do **saldo médio** indicam que alguns clientes têm saldos consideravelmente altos, possivelmente sugerindo maior risco de inadimplência nesses casos.

A **renda anual** dos clientes tem uma grande dispersão, o que indica que o banco está atendendo tanto clientes de baixa quanto de alta renda. A diferença entre o valor mínimo e o máximo e o alto desvio padrão sugerem que existem subgrupos de clientes com perfis financeiros muito distintos.





Também é possível verificar que não há valores ausentes em cada coluna. Sendo assim, não é necessário realizar um tratamento mais aprofundado desses dados.

```
Valores ausentes em cada coluna:
target          0
estudante       0
saldo médio     0
renda anual     0
dtype: int64
```

Após a implementação do modelo de regressão logística do Scikit-Learn e a separação dos dados de teste e treino em 30% e 70% respectivamente. O modelo foi treinado usando cada uma das variáveis separadamente como preditoras e depois com todas juntas.

Interpretando os coeficientes do modelo, temos o impacto das principais variáveis (saldo médio, renda anual e estudante).

	Variável	Coeficiente
0	saldo médio	0.005674
1	estudante	-0.645219
2	renda anual	0.000004

O coeficiente de `saldo médio` indica que para cada aumento de uma unidade no saldo médio, a razão de chances (odds ratio) de inadimplência aumenta em cerca de 0,57%. Isso significa que clientes com um saldo médio mais alto tendem a ter uma probabilidade ligeiramente maior de serem inadimplentes.

O coeficiente negativo ligado a variável categórica `estudante` indica que a razão de chances de inadimplência diminui em relação aos não estudantes. Em resumo, nosso modelo nos diz que ser estudante está associado a uma menor probabilidade

de inadimplência. Indicando algum comportamento desse grupo que demonstra uma abordagem mais cautelosa ao lidar com crédito.

O coeficiente de `renda anual`, apesar de muito baixo, indica um pequeno impacto na probabilidade de inadimplência quando outras variáveis são levadas em conta.

Nosso modelo apresentou as seguintes acurácias relacionadas às variáveis sozinhas e em conjunto.

```
Acurácia usando 'saldo médio' como preditor: 0.8785
Acurácia usando 'estudante' como preditor: 0.5468
Acurácia usando 'renda anual' como preditor: 0.5423
Acurácia usando todas as variáveis como preditoras: 0.8814
```

Sabendo que,

$$\text{Acurácia} = \frac{TP + TN}{TP + TN + FP + FN}$$

que

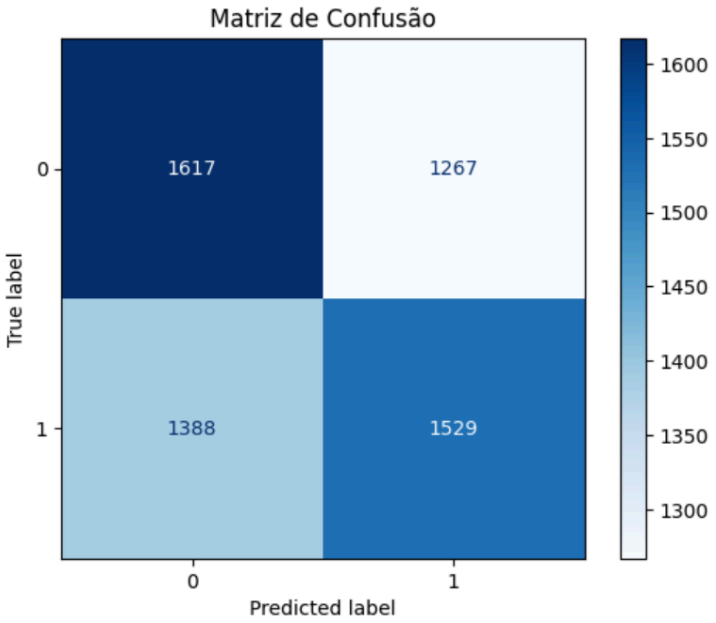
$$\text{Precisão} = \frac{TP}{TP + FP}$$

e que

$$\text{Recall} = \frac{TP}{TP + FN}$$

E observando nossa matriz de confusão,

	Previsto: Não Inadimplente	**Previsto: Inadimplente**
Real: Não Inadimplente	Verdadeiros Negativos (TN)	Falsos Positivos (FP)
Real: Inadimplente	Falsos Negativos (FN)	Verdadeiros Positivos (TP)



$$\text{Acurácia} = \frac{TP + TN}{TP + TN + FP + FN} = \frac{1617 + 1529}{1617 + 1529 + 1267 + 1388} \approx 0.6$$

A acurácia geral de aproximadamente 63% sugere que, apesar de o modelo ter conseguido prever corretamente um pouco mais da metade dos casos, ele ainda tem uma quantidade considerável de erros.

$$\text{Precisão} = \frac{TP}{TP + FP} = \frac{1529}{1529 + 1267} \approx 0.55$$

A precisão de 55% significa que, quando o modelo prevê que um cliente será inadimplente, ele está certo em 55% das vezes. O número de **falsos positivos** é alto, o que indica que o modelo marca erroneamente muitos clientes como inadimplentes.

$$\text{Recall} = \frac{TP}{TP + FN} = \frac{1529}{1529 + 1388} \approx 0.52$$

O recall de 52% significa que o modelo detecta corretamente 52% dos inadimplentes. Isso é relativamente baixo, sugerindo que o modelo perde quase metade dos clientes que realmente se tornam inadimplentes.

Um método para tornar nosso modelo mais sensível à inadimplêndia pode ser o ajuste do limiar. O ajuste do limiar padrão (0.5) para 0.3 faz com que nosso modelo preveja casos de inadimplência se as chances calculadas (Odds) for maior que 0.3.

`Métricas com limiar de decisão 0.3:`

```
Acurácia: 0.8607
Precisão: 0.8082
Revocação (Recall): 0.9479
F1 Score: 0.8725
Matriz de Confusão:
[[2228  656]
 [ 152 2765]]
```

É visível o impacto gerado em todas as métricas observadas. O ajuste para um limiar menor torna o modelo mais "agressivo" na detecção de inadimplência. Esse comportamento é favorável em situações onde perder um inadimplente (falso negativo) é mais custoso do que rotular um cliente não inadimplente como inadimplente (falso positivo).

Conclusões Gerais

- **Saldo médio é o melhor preditor de inadimplência**, o que faz sentido, pois clientes com saldos mais altos podem ter maiores dificuldades de pagamento.
- **As variáveis ‘estudante’ e ‘renda anual’ não são bons preditores isoladamente** e inicialmente parecem não acrescentar muito ao modelo quando consideradas sozinhas.
- **Com todas as variáveis**, o modelo atinge uma acurácia de 88.14%, mas a matriz de confusão ainda mostra uma quantidade significativa de erros, especialmente falsos positivos e falsos negativos. Isso pode indicar que o modelo ainda não é ideal para detectar todos os casos de inadimplência com precisão.